Organization TC1600 Bldg./Room REMSEN

U. S. DEPARTMENT OF COMMERCE
COMMISSIONER FOR PATENTS
P.O. BOX 1450
ALEXANDRIA, VA 22313-1450
IF UNDELIVERABLE RETURN IN TEN DAYS

OFFICIAL BUSINESS

U.S. OFFICIAL MAIL
PENALTY FOR PRIVATE USE $300

02 1A $ 01.98
0004204034 FEB 08 200
MAILED FROM ZIPCODE 2231

AN EQUAL OPPORTUNITY EMPLOYER

RETURNED TO SENDER
FORWARDING ORDER EXPIRED

RECEIVED

FEB 18 2005

TECH CENTER 1600/2900

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
10/063,537	05/02/2002	Audrey Goddard	P3230R1C001-168	1051

7590 02/08/2005

ANNE MARIE KAISER
KNOBBE MARTENS OLSON & BEAR
620 NEWPORT CENTER DRIVE
NEWPORT BEACH, CA 92660

EXAMINER	
SEHARASEYON, JEGATHEESAN	
ART UNIT	**PAPER NUMBER**
1647	

DATE MAILED: 02/08/2005

Please find below and/or attached an Office communication concerning this application or proceeding.

Office Action Summary	Application No.	Applicant(s)	
	10/063,537	EATON ET AL.	
	Examiner	Art Unit	
	Jegatheesan Seharaseyon	1647	

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address --

Period for Reply

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE 3 MONTH(S) FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If the period for reply specified above is less than thirty (30) days, a reply within the statutory minimum of thirty (30) days will be considered timely.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

Status

1)☒ Responsive to communication(s) filed on *16 September 2004*.

2a)☒ This action is **FINAL**. 2b)☐ This action is non-final.

3)☐ Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

Disposition of Claims

4)☒ Claim(s) *1-5* is/are pending in the application.

4a) Of the above claim(s) _____ is/are withdrawn from consideration.

5)☐ Claim(s) _____ is/are allowed.

6)☒ Claim(s) *1-5* is/are rejected.

7)☐ Claim(s) _____ is/are objected to.

8)☐ Claim(s) _____ are subject to restriction and/or election requirement.

Application Papers

9)☐ The specification is objected to by the Examiner.

10)☒ The drawing(s) filed on *02 May 2002* is/are: a)☐ accepted or b)☐ objected to by the Examiner.

Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).

Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).

11)☐ The oath or declaration is objected to by the Examiner. Note the attached Office Action or form PTO-152.

Priority under 35 U.S.C. § 119

12)☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).

a)☐ All b)☐ Some * c)☐ None of:

1.☐ Certified copies of the priority documents have been received.

2.☐ Certified copies of the priority documents have been received in Application No. _____.

3.☐ Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).

* See the attached detailed Office action for a list of the certified copies not received.

Attachment(s)

1) ☒ Notice of References Cited (PTO-892)
2) ☐ Notice of Draftsperson's Patent Drawing Review (PTO-948)
3) ☒ Information Disclosure Statement(s) (PTO-1449 or PTO/SB/08) Paper No(s)/Mail Date *9/16/2004*.
4) ☐ Interview Summary (PTO-413) Paper No(s)/Mail Date. _____ .
5) ☐ Notice of Informal Patent Application (PTO-152)
6) ☐ Other: _____.

DETAILED ACTION

1. The Amendment and Declarations under 37 CFR § 1.132, both submitted 16 September 2004, have been entered. Claim 1 is amended. Claim 6 is cancelled. Claims 1-5 is under examination in the Instant Application.

2. The text of those sections of Title 35, U.S. Code, not included in this action can be found in a prior Office action.

3. The Office acknowledges the receipt of the drawings on 5/2/2002.

4. The Office acknowledges the change in title.

5. The Office also acknowledges the removal of embedded hyperlinks.

6. The Applicants have provided a copy of the sequence listing in response to the "Notice to Comply".

7.The request for the deletion of an inventor in this nonprovisional application under 37 CFR 1.48(b) is deficient because: The request was not accompanied by the statement required under 37 CFR 1.48(b)(2). Applicants are required to state that the deletion is required because claims have been amended or canceled such that he or she is no longer an inventor of any remaining claim in the non-provisional application.

8. The Office acknowledges the submission of the IDS dated 9/16/2004.

9. All pending rejections of claim 6 are withdrawn because Applicants have cancelled claim 6.

Priority

10. Applicant has not complied with one or more conditions for receiving the benefit of an earlier filing date under 35 U.S.C. 119. Applicants have argued that they are

entitled to the benefit of the filing date of August 24, 2000 based on the disclosure amplification data in PCT/US00/23328. Although, the previous patent application discloses the same polypeptide sequences (SEQ ID NO: 31 and 32) as the instant specification, the disclosure is not enabling for the instant invention and therefore do not impart utility to the claims of the current application. Therefore, the filing date of 2 May 2002 is maintained as the priority date.

35 USC § 112, second paragraph, withdrawn

11. The rejection of claims 1-6 under 35 USC 112, second paragraph for being vague and indefinite, as set forth in the Office Action of 15 June 2004 is withdrawn is response to Applicants amendments and arguments.

35 U.S.C. § 101/112, first paragraph, Lack of Utility, Enablement, maintained

12. Claims 1-5 are rejected under 35 U.S.C. 101, as lacking utility. The reasons for this rejection under *35 U.S.C. § 101* are set forth at pp. 3-8 of the previous Office Action (15 June 2004). Claims 1-5 are also rejected under 35 U.S.C. 112, first paragraph. Specifically, since the claimed invention is not supported by either a specific and substantial asserted utility or a well established utility for the reasons set forth in the previous Office Action (15 June 2004), one skilled in the art clearly would not know how to use the claimed invention.

Applicants argue (16 September 2004, page 6) that the results presented in the instant specification are enabling for the polypeptide of SEQ ID NO: 32 and antibodies directed against polypeptide. They argue that the utilities of PRO1115 polypeptide

include the use as a diagnostic tool, as well as therapeutically as a target for treatment, based on the data that PRO1115 cDNA is more highly expressed normal stomach tissue or normal lung compared to stomach tumor tissue or lung tumor tissue. Applicants has also extensively discussed the utility guidelines (pages 6-8). Applicant's arguments (16 September 2004) have been fully considered but are not found to be persuasive for the following reasons:

In the instant case, the specification provides data showing that polynucleotide is more highly expressed in normal stomach or lung compared to stomach or lung tumor tissue. However, there is no further supporting evidence to indicate that the polypeptide encoded by the polynucleotide of the instant invention is also differentially expressed in the normal tissue compared to the tumor tissue and as such one of skill in the art would conclude that it is not supported by a substantial asserted utility or a well-established utility. Furthermore, as discussed extensively by Pennica et al. in the previous Office Action (15 June 2004, page 7), what is often seen is a *lack* of correlation between DNA amplification and increased gene expression. Although, the Office, in the Office Action mailed on the 15 June 2004 provided evidence taught by Sen (page 6) that cancerous tissue is known to be aneuploid, and thus, a higher amplification of a gene does not necessarily mean higher expression in that tissue, but can merely be an indication that the tissue in question is aneuploid, the Applicants assert that they fail to see how it is relevant to the utility of the disclosed nucleic acids, or their corresponding polypeptides whether the differential expression reported in example 18, is due to aneuploidy or not. The relevance of this teachings are associated with the abnormal numbers of

chromosome often present in the cancerous tissue and the lack of correction in the instant invention for aneuploidy.

As discussed by Haynes et al (1998, Electrophoresis, 19: 1862-1871), polypeptide levels cannot be accurately predicted from mRNA levels, and that, according to their results, the ratio varies from zero to 50-fold (page 1863). The literature cautions researchers against drawing conclusions based on small changes in transcript expression levels between normal and cancerous tissue. For example, Hu et al. (2003, Journal of Proteome Research 2: 405-412) analyzed 2286 genes that showed a greater than 1-fold difference in mean expression level between breast cancer samples and normal samples in a microarray (p. 408, middle of right column). Hu et al. discovered that, for genes displaying a 5-fold change or less in tumors compared to normal, there was no evidence of a correlation between altered gene expression and a known role in the disease. However, among genes with a 10-fold or more change in expression level, there was a strong and significant correlation between expression level and a published role in the disease (see discussion section).

Given the increase in amplified DNA (DNA copy number) for PRO1115 in the in normal stomach or lung compared to stomach or lung tumor tissue, and the evidence provided by the current literature, it is clear that one skilled in the art would not assume that a higher expression would correlate with increased mRNA or polypeptide levels. Further research needs to be done to determine whether the decrease of PRO1115 DNA compared normal stomach or lung tissues supports a role for the peptide in the cancerous tissue; such a role has not been suggested by the instant disclosure. Such

further research requirements make it clear that the asserted utility is not yet in currently available form, i.e., it is not substantial. This further experimentation is part of the act of invention and until it has been undertaken, Applicant's claimed invention is incomplete. As discussed in Brenner v. Manson, (1966, 383 U.S. 519, 148 USPQ 689), the court held that:

> "The basic quid pro quo contemplated by the Constitution and the Congress for granting a patent monopoly is the benefit derived by the public from an invention with substantial utility", "[u]nless and until a process is refined and developed to this point-where specific benefit exists in currently available form-there is insufficient justification for permitting an applicant to engross what may prove to be a broad field", and,
>
> "a patent is not a hunting license", "[i]t is not a reward for the search, but compensation for its successful conclusion."

Accordingly, the Specification's assertions that the claimed PRO1115 polypeptides have utility in the fields of cancer diagnostics and cancer therapeutics are not substantial.

The declarations of Mr. Grimaldi, filed under 37 CFR 1.132 (16 September 2004), is insufficient to overcome the rejection of claims 1-5, based upon 35 U.S.C. § 101 and 35 U.S.C. § 112, first paragraph as set forth in the last Office action. Similarly, the declaration of Dr. Polakis, filed under 37 CFR 1.132 (16 September 2004), is insufficient to overcome the rejection of claims 1-5, based upon 35 U.S.C. § 101 and 35 U.S.C. § 112, first paragraph as set forth in the last Office action. Likewise, the declaration of Dr. Ashkenazi, filed under 37 CFR 1.132 (16 September 2004), is insufficient to overcome the rejection of claims 1-5, based upon 35 U.S.C. § 101 and 35 U.S.C. § 112, first paragraph as set forth in the last Office action because:

In the declaration filed under37 CFR 1.132 (16 September 2004, originally filed in application serial number 10/063,557), senior research associate Mr. Grimaldi states (page 2, paragraph 5), that "data from pooled samples is more likely to be accurate than data obtained from a sample from a single individual". In addition, Mr. Grimaldi declaration on paragraphs 6 and 7 states that semi-quantitative analysis employed to generate the data of example 18 is sufficient to determine if a gene is over or under expressed in tumor cells compared to corresponding normal tissue. Further it asserted that that any visually detectable difference seen between two samples is indicative of at least a two-fold difference in cDNA between the tumor tissue and the counterpart normal tissue. Mr. Grimaldi also asserted that, if a difference is detected, this indicates that the gene and its corresponding polypeptide and antibodies against the polypeptide are useful for diagnostic purposes, to screen samples to differentiate between normal and tumor. It is further stated that additional studies can then be conducted if further information is desired. In paragraph 7, declarant indicates that the difference in the expression is expected to be reflected in the difference in the corresponding protein. However, this appears to be declarant's opinion, and is not supported by fact or evidence and there has been no distinction on the record in general or in the specification as filed between total nucleic acid, which includes chromosomal DNA, and mRNA. There is no description in the specification to that would indicate a correlation with higher or lower expression levels of the message to the PRO1115. It remains that, there is no information on the record as to whether the claimed protein is expressed at all in the stomach and lung tissue, cancerous or otherwise. Furthermore, it remains that,

as evidenced by Pennica et al., the issue is simply not predictable, and the specification presents a mere invitation to experiment.

Applicants citing the second Grimaldi declaration (exhibit 2) filed under 37 CFR § 1.132 argues that, "Those who work in this field are well aware that in the vast majority of cases, when a gene is over-expressed..... this same principal applies to gene under-expression." Again citing paragraph 5, Applicants contend that 'the detection of increased mRNA expression is expected to result in increased polypeptide expression, and detection of decreased mRNA expression is expected to result in decreased polypeptide expression. The detection of increased or decreased polypeptide expression can be used for the diagnosis and treatment."

The Polakis Declaration states that approximately 200 gene transcripts were identified that are present in human tumor cells at significantly higher levels than in control tissues and that antibodies have been developed that identify and could possibly be used to down regulate the PRO peptides. Dr. Polakis states that it remains a central dogma in molecular biology that increased mRNA levels are predictive of corresponding increased levels of the encoded polypeptide. Dr Polakis characterizes the instances where such a correlation does not exist as exceptions to the rule. Only Dr. Polakis conclusions are provided in the declaration. There is no evidentiary support to Dr. Polakis' statement that it remains a central dogma in molecular biology that increased mRNA Levels are predictive of corresponding increased levels of the encoded polypeptide.

Applicants also refer to three additional articles (Orntoft et al., Hyman et al., and Pollack et al. as providing evidence that gene amplification generally results in elevated levels of the encoded polypeptide. Applicants characterize Orntoft et al. as teaching in general (18 of 23 cases) chromosomal areas with more than 2-fold gain of DNA showed corresponding increase in mRNA transcripts. Applicants further characterize Hyman et al. as providing evidence of a prominent global influence of copy number changes on gene expression levels. It is also claimed by the Applicants that Pollack et al. teach that 62% of highly amplified genes show moderately or highly elevated expression and that, on average, a 2-fold change in DNA copy number is associated with a 1.5-fold change in mRNA levels.

Orntoft et al. appear to have looked at increased DNA content over large regions of chromosomes and comparing that to mRNA and polypeptide levels from the chromosomal region. Their approach to investigating gene copy number was termed CGH. Orntoft et al. do not appear to look at gene amplification, mRNA levels and polypeptide levels from a single gene at a time. The instant specification reports data regarding amplification of individual gene, which may or may not be in a chromosomal region, which is highly amplified. Orntoft et al. concentrated on regions of chromosomes with strong gains of chromosomal material containing clusters of genes (p.40). This analysis was not done for PRO1115 in the instant specification. That is, it is not clear whether or not PRO1115 is in a gene cluster in a region of a chromosome that is highly amplified. Therefore, the relevance, if any of Orntoft et al. is not clear. Hyman et al. also used CGH approach in their research. Less than half (44%) of highly amplified genes

showed over expression (abstract). Polypeptide levels were not investigated. Therefore, Hyman et al. also do not support utility of the polypeptides of the instant invention. Pollack et al. using CGH technology, concentrate on large chromosome regions showing high amplification (p.12965). However, Pollack et al. did not investigate or show a relationship with amplification and polypeptide expression. In fact the authors caution that elevated expression of an amplified gene cannot alone be considered strong independent evidence of candidate oncogene's role in tumorigenesis. Thus, these references collectively do not teach as Applicants contend that there is a direct correlation between increased mRNA levels and increased levels of encoded protein. Accordingly, the Applicants' assertions that the PRO1115 polypeptides have utility in the cancer diagnostics are not substantial.

Applicants also contend that the claimed antibodies would have diagnostic utility even if there is no positive correlation between gene expression and expression of the encoded polypeptide. Further, it is asserted that even if there was no correlation between gene expression and increased or decreased protein expression for PRO1115, the polypeptide encoded by a gene that is over-expressed or under expressed in cancer would still have credible, specific and substantial utility. Applicants assert that this position is supported by the declaration filed under 37 CFR 1.132 (16 September 2004) by staff scientist Ashkenazi. It claims that the purpose of the experiments that measured increases in gene copy number was to identify tumor cell markers useful for cancer treatment (pages 1-2, Declaration, 16 September 2004) and to identify cancers for which there was an absence of gene product over-expression (page 2). The Ashkenazi

declaration further argues that, even when amplification of a gene in a tumor does not correlate with an increase in polypeptide expression, the absence of the gene product over-expression still provides significant information for cancer diagnosis and treatment.

Applicants argue (Response, 16 September 2004, page 14) that even if a prima facie case of lack of utility has been established, it should be withdrawn on consideration of the totality of the evidence. Applicants provide evidence in the form of a publication by Hanna et al. (attached to the Response of 16 September 2004). Applicants contend that the publication teaches that the HER-2/neu gene is over-expressed in breast cancers, and teaches that diagnosis of breast cancer includes testing both the amplification of the HER-2/neu gene as well as over- expression of the HER-2/neu gene product. Applicant argues that the disclosed assay leads to a more accurate classification of the cancer and a more effective treatment of it. The examiner agrees. In fact, Hanna et al. supports the instant rejection, in that Hanna et al. show that gene amplification does not reliably correlate with polypeptide over-expression, and thus the level of polypeptide expression must be tested empirically.

Applicants' arguments and declarations have been fully considered but are deemed not to be persuasive. In the instant application gene expression studies were conducted using pooled samples of normal and tumor tissues. With reference to Grimaldi reference, this appears to be declarant's own opinion, and is not supported by fact or evidence. In addition, one cannot determine from the data in the specification whether the observed "amplification" of nucleic acid is due to increase in copy number, or alternatively due to increase in transcription rates. It is important to note that the

instant specification provides no information regarding differential mRNA levels of PRO1115 in normal stomach or normal lung compared to stomach tumor or lung tumor tissue samples. The specification describes only gene amplification data. The argument presented evinces that instant specification provides a mere invitation to experiment, and not readily available utility. The declaration does not provide data such that the examiner can independently draw conclusions. In addition, there is no evidentiary art that would corroborate for example, that "any visually detectable difference seen between two samples is indicative of at least a two-fold difference in cDNA between the tumor tissue and the counterpart normal tissue." Furthermore, as indicated above the literature cautions researchers against drawing conclusions based on small changes in transcript expression levels between normal and cancerous tissue (see Haynes et al. and Hu et al discussions above). It is also not known whether PRO1115 polypeptide is expressed in normal stomach and lung tissue and what the relative levels of expression are. In the absence of any of the above information, all that the specification does is present evidence that the DNA encoding PRO1115 is amplified at higher levels in normal stomach or normal lung compared to stomach tumor and lung tumor tissues, and invite the artisan to determine the rest of the story. This is further borne out by Grimaldi assertion that "additional studies can then be conducted if further information is desired" (Appendix 1, paragraph 7). Such is insufficient to meet the requirements of 35 U.S.C. § 101 utility for the claimed protein.

Although, Dr. Polakis states that it remains a central dogma in molecular biology that increased mRNA levels are predictive of corresponding increased levels of the

encoded polypeptide, it is important to note that the instant specification provides no information regarding differential mRNA levels of PRO1115 in tumor samples as contrasted to normal tissue samples or the corresponding protein levels. Only gene amplification data were presented. Therefore, the declaration is insufficient to overcome the rejection of claims 1-5 based upon 35 U.S.C. § 101 and 112, first paragraph, since it is limited to a discussion of data regarding the correlation of mRNA levels and polypeptide levels. Furthermore, the declarations do not provide data such that the examiner can independently draw conclusions. Finally, it is noted that the literature cautions researchers from drawing conclusions based on small changes in transcript expression levels between normal and cancerous tissue. For example, as discussed above, Hu et al. (2003, Journal of Proteome Research 2:405-412) analyzed 2286 genes that showed a greater than 1-fold difference in mean expression level between breast cancer samples and normal samples in a microarray (p. 408, middle of right column) and discovered that, for genes displaying a 5-fold change or less in tumors compared to normal, there was no evidence of a correlation between altered gene expression and a known role in the disease. However, among genes with a 10-fold or more change in expression level, there was a strong and significant correlation between expression level and a published role in the disease (see discussion section).

The declaration of Ashkenazi appears to argue that even if there was no correlation between gene expression and increased or decreased protein expression for PRO1115, the polypeptide encoded by a gene that is over-expressed or under expression in cancer would still have credible, specific and substantial utility. The

examiner agrees that evidence regarding lack of over-expression would be useful. However, there is no evidence as to whether the gene products (such as the polypeptide) are over-expressed or not. Further research is required to determine such. Thus, the asserted utility is not substantial.

Although, Applicants agree that Sen reference teaches that most cancerous tissues are aneuploid, it is argued on page 10, 3rd paragraph of remarks that there is no relevance to the expression levels and to the state of aneuploidy of the tumor cells with respect to the asserted utility. Applicants claim regardless of the cause of the differential expression, the fact that there is a higher level or lower level of expression of PRO1115 gene in normal stomach or lung tissue compared to tumor containing stomach and lung tissue allows this gene expression to be used as a diagnostic tool. These arguments have been fully considered but are not found to be persuasive because as indicated in the Office Action of 15 June 2004, the differential expression can merely be an indication that the cancer tissue is aneuploid (see page 7 of the office Action). In addition, the lack of information on the record whether the claimed protein (PRO1115) is expressed at all stomach and lung tissue, cancerous or otherwise would make significant further research a necessity.

At page 10, Applicants assert that they have established that the accepted understanding in the art is that there is a direct correlation between mRNA levels and the level of expression of the encoded protein. It is also asserted that the Office relying on Pennica et al reference is also stating that data pertaining to PRO1115 polynucleotides do not necessarily indicate anything significant regarding the claimed

PRO1115 polypeptides. Applicants further assert that the office is confusing the relationship between an increase in copy number of a gene or gene amplification on the one hand, and increased expression of a gene or mRNA expression on the other. These arguments have been fully considered but are not found to be persuasive. Haynes et al. and Hu et al. teachings listed above discussed above contradict Applicants assertion that there exists a direct correlation between mRNA levels and the level of expression of the encoded protein. In fact, the literature cautions researchers against drawing conclusions based on small changes in transcript expression levels between normal and cancerous tissues. The Office relied on Pennica et al. to teach that, "it does not necessarily follow that an increase in gene copy number results in increased gene expression". Pennica et al. on p.14722, clearly discuss the variability in DNA amplification and gene expression. Contrary to Applicants assertion that "it is possible that the apparent amplification observed for *WISP-2* may be caused by another gene in this amplicon" (see bottom of p.10) further reading of Pennica et al. indicates that the reduced expression of *WISP-2* in colon tumors and cell lines suggests that it may function as a tumor suppressor. Finally, with respect to Applicants assertion that the Office is confusing the relationship between an increase in copy number of a gene or gene amplification on the one hand, and increased expression of a gene or mRNA expression on the other, it is the position of the Office that there is no confusion with respect to the lack of correlation of between DNA amplification and gene expression (see p.14722, left column).

The Office agrees with the Applicants in that the Pennica et al. reference does not discuss the relationship of level of mRNA and level of protein expression. However, this reference was cited by the Office to show the lack of correlation of between DNA amplification and gene expression. Although, Applicants indicate on p.10 that there is a well established correlation in the art that the level of protein is positively correlated to the level of mRNA, as indicated above Haynes et al. and Hu et al., polypeptide levels cannot be accurately predicted from mRNA levels. Therefore, there is no evidence to support Applicants' assertion that there is working hypothesis among those skilled in the art is that there is a direct correlation between mRNA levels and protein levels. In addition, even if there was a correlation between mRNA levels and protein levels, Applicants have not established a nexus between the DNA of instant invention and PRO1115 protein. As stated above and in the Office Action of 15 June 2004, the specification does not provide sufficient evidence or guidance to the skilled artisan to diagnose or treat any disease. Therefore, there would be no specific utility for antibodies of PRO1115 protein.

Therefore, all of these reasons, the rejection of claims 1-5 based upon 35 U.S.C. § 101 and 35 U.S.C. § 112, first paragraph as set forth in the last Office Action is maintained.

Claim Rejections - 35 USC § 103, maintained

13. The rejection of claims 1-5 under 35 U.S.C. 103(a) as being unpatentable over Collier et al. (Accession No. Q9BWY7, June 2001) in view of Turner et al. (Accession No. AAX146414, WO134804A1, published May 2001) is maintained because as

indicated above in paragraph 10, the instant application has been denied the earlier priority date. In addition Applicants declaration submitted under 37 CFR 1.131 has been considered but not persuasive because the disclosure is not enabling as indicated in paragraph 9 above. Therefore, teachings of Collier et al. and Turner et al. are considered prior art and the rejection is maintained.

14. No claims are allowed.

15. **THIS ACTION IS MADE FINAL.** See MPEP § 706.07(a). Applicant is reminded of the extension of time policy as set forth in 37 CFR 1.136(a).

A shortened statutory period for reply to this final action is set to expire THREE MONTHS from the mailing date of this action. In the event a first reply is filed within TWO MONTHS of the mailing date of this final action and the advisory action is not mailed until after the end of the THREE-MONTH shortened statutory period, then the shortened statutory period will expire on the date the advisory action is mailed, and any extension fee pursuant to 37 CFR 1.136(a) will be calculated from the mailing date of the advisory action. In no event, however, will the statutory period for reply expire later than SIX MONTHS from the date of this final action.

Contact Information

Any inquiry concerning this communication or earlier communications from the examiner should be directed to Jegatheesan Seharaseyon whose telephone number is 571-272-0892. The examiner can normally be reached on M-F: 8:30-4:30.

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Brenda Brumback can be reached on 571-272-0961. The fax phone number for the organization where this application or proceeding is assigned is 703-872-9306. Information regarding the status of an application may be obtained from the Patent Application Information Retrieval (PAIR) system. Status information for published applications may be obtained from either Private PAIR or Public PAIR. Status information for unpublished applications is available through Private PAIR only. For more information about the PAIR system, see http://pair-direct.uspto.gov. Should you have questions on access to the Private PAIR system, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free).

JS 1/05

Brenda Brumback

BRENDA BRUMBACK
SUPERVISORY PATENT EXAMINER
TECHNOLOGY CENTER 1600

OIPE
SEP 16 2004
PATENT & TRADEMARK OFFICE

FORM PTO-1449 U.S. DEPARTMENT OF COMMERCE
PATENT AND TRADEMARK OFFICE

SUPPLEMENTAL
INFORMATION DISCLOSURE STATEMENT
BY APPLICANT

(USE SEVERAL SHEETS IF NECESSARY)

ATTY. DOCKET NO.	APPLICATION NO.
GNE.3230R1C25	10/083,537
APPLICANT: Goddard, et al. (as amended)	
FILING DATE: May 2, 2002	GROUP: 1647

U.S. PATENT DOCUMENTS

EXAMINER INITIAL		DOCUMENT NUMBER	DATE	NAME	CLASS	SUBCLASS	FILING DATE (IF APPROPRIATE)
SS	1	5,536,637	07/16/96	Jacobs			

FOREIGN PATENT DOCUMENTS

EXAMINER INITIAL		DOCUMENT NUMBER	DATE	COUNTRY	CLASS	SUBCLASS	TRANSLATION YES	TRANSLATION NO

EXAMINER INITIAL		OTHER DOCUMENTS (INCLUDING AUTHOR, TITLE, DATE, PERTINENT PAGES, ETC.)
SS	2	Database Search, DNA Sequence Alignments [BLASTN 2.2.1 [Jul-12-2001], NCBI]
SS	3	Database Search, DNA Search, Protein Sequence Alignments [BLASTP 2.2.1 [Jul-12-2001], NCBI]

S:\DOCS\DOH\DOH-8072.DOC:dmb
081904

EXAMINER	DATE CONSIDERED
	2/26/05

*EXAMINER: INITIAL IF CITATION CONSIDERED, WHETHER OR NOT CITATION IS IN CONFORMANCE WITH MPEP 609; DRAW LINE THROUGH CITATION IF NOT IN CONFORMANCE AND NOT CONSIDERED, INCLUDE COPY OF THIS FORM WITH NEXT COMMUNICATION TO APPLICANT.

Notice of References Cited	Application/Control No. 10/063,537	Applicant(s)/Patent Under Reexamination EATON ET AL.	
	Examiner Jegatheesan Seharaseyon	Art Unit 1647	Page 1 of 1

U.S. PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Name	Classification
	A	US-			
	B	US-			
	C	US-			
	D	US-			
	E	US-			
	F	US-			
	G	US-			
	H	US-			
	I	US-			
	J	US-			
	K	US-			
	L	US-			
	M	US-			

FOREIGN PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Country	Name	Classification
	N					
	O					
	P					
	Q					
	R					
	S					
	T					

NON-PATENT DOCUMENTS

*		Include as applicable: Author, Title Date, Publisher, Edition or Volume, Pertinent Pages)
	U	Haynes et al., Proteome analysis: Biological Assay or data archive? (1998), Electophoresis, Vol.19, pp1862-1871.
	V	Hu et al., Analysis of genomic and Proteomic Data using Advanced literature mining (2003), J. of Proteome Research, Vol.2, pp405-412.
	W	
	X	

*A copy of this reference is not being furnished with this Office action. (See MPEP § 707.05(a).)
Dates in MM-YYYY format are publication dates. Classifications may be US or foreign.

research articles Journal of proteome research

Analysis of Genomic and Proteomic Data Using Advanced Literature Mining

Yanhui Hu, Lisa M. Hines, Haifeng Weng, Dongmei Zuo, Miguel Rivera, Andrea Richardson, and Joshua LaBaer*

Institute of Proteomics, Harvard Medical School—BCMP, 240 Longwood Avenue, Boston, Massachusetts 02115

Received March 13, 2003

High-throughput technologies, such as proteomic screening and DNA micro-arrays, produce vast amounts of data requiring comprehensive analytical methods to decipher the biologically relevant results. One approach would be to manually search the biomedical literature; however, this would be an arduous task. We developed an automated literature-mining tool, termed MedGene, which comprehensively summarizes and estimates the relative strengths of all human gene−disease relationships in Medline. Using MedGene, we analyzed a novel micro-array expression dataset comparing breast cancer and normal breast tissue in the context of existing knowledge. We found no correlation between the strength of the literature association and the magnitude of the difference in expression level when considering changes as high as 5-fold; however, a significant correlation was observed ($r = 0.41$; $p = 0.05$) among genes showing an expression difference of 10-fold or more. Interestingly, this only held true for estrogen receptor (ER) positive tumors, not ER negative. MedGene identified a set of relatively understudied, yet highly expressed genes in ER negative tumors worthy of further examination.

Keywords: bioinformatics • micro-array • text mining • gene-disease association • breast cancer

Introduction

At its current pace, the accumulation of biomedical literature outpaces the ability of most researchers and clinicians to stay abreast of their own immediate fields, let alone cover a broader range of topics. For example, to follow a single disease, e.g., breast cancer, a researcher would have had to scan 130 different journals and read 27 papers per day in 1999.[1] This problem is accentuated with high-throughput technologies such as DNA micro-arrays and proteomics, which require the analysis of large datasets involving thousands of genes, many of which are unfamiliar to a particular researcher. In any microarray experiment, thousands of genes may demonstrate statistically significant expression changes, but only a fraction of these may be relevant to the study. The ability to interpret these datasets would be enhanced if they could be compared to a comprehensive summary of what is known about all genes. Thus, there is a need to summarize existing knowledge in a format that allows for the rapid analysis of associations between genes and diseases or other specific biological concepts.

One solution to this problem is to compile structured digital resources, such as the Breast Cancer Gene Database[1] and the Tumor Gene Database.[2] However, as these resources are hand-curated, the labor-intensive review process becomes a rate-limiting step in the growth of the database. As a result, these databases have a limited scale and the genes are not selected in a systematic fashion.

An alternative approach is automated text mining; a method which involves automated information extraction by searching documents for text strings and analyzing their frequency and context. This approach has been used successfully in several instances for biological applications. In most cases, it has been applied to extract information about the relationships or interactions that proteins or genes have with one another, in the literature or by functional annotation.[3−7] Thus far, few publication have applied text-mining to examine the global relationships between genes and diseases. Perez-Iratxeta et al. automatically examined the GO (Gene Ontology) annotation of genes and their predicted chromosomal locations in order to identify genes linked to inherited disorders.[8]

To obtain a more global understanding of disease development, it would be valuable to incorporate information regarding all possible gene-disease relationships, including biochemical, physiological, pharmacological, epidemiological, as well as genetic. This information would enable comprehensive comparisons between large experimental datasets and existing knowledge in the literature. This would accomplish two things. First, it would serve to validate experiments by demonstrating that known responses occur as predicted. Second, it would rapidly highlight which genes are corroborated by the literature and which genes are novel in a given context. We have utilized a computational approach to literature mining to produce a

* To whom correspondence should be addressed: jlabaer@hms.harvard.edu.

10.1021/pr0340227 CCC: $25.00 © 2003 American Chemical Society

Published on Web 06/13/2003

comprehensive set of gene-disease relationships. In addition, we have developed a novel approach to assess the strength of each association based on the frequency of citation and co-citation. We applied this tool to help interpret the data from a large micro-array gene expression experiment comparing normal and cancerous breast tissue.

Methods

MedGene Database. MedGene is a relational database, storing disease and gene information from NCBI, text mining results, statistical scores, and hyperlinks to the primary literature. MedGene has a web-based user interface for users to query the database (http://hipseq.med.harvard.edu/MedGene/).

Text Mining Algorithms. MeSH files were downloaded from the MeSH web site at NLM (Nation Library of Medicine) (http://www.nlm.nih.gov/mesh/meshhome.html) and human disease categories were selected. LocusLink files were downloaded from the LocusLink web site at NCBI (http://www.ncbi.nih.gov/LocusLink/). Official/preferred gene symbol, official/preferred gene name, and gene alternative symbols and names, all relevant annotations and URLs for each LocusLink record, were collected. Gene search terms were used for literature searching and included all qualified gene names, gene symbols, and gene family terms. Primary gene keys, predominantly qualified gene family terms and gene official/preferred symbols, were used to index Medline records. If the official/preferred gene symbols did not meet the standards to be an index, then qualified gene official/preferred names were used. A local copy of Medline records (up to July, 2002) was pre-selected.

A JAVA module examined the MeSH terms and then indexed each Medline record with the appropriate disease terms. A separate JAVA module was used to examine the titles and abstracts for gene search terms and then to index the gene-related Medline records with the relevant primary gene key(s).

Statistical Methods. For every gene and disease pair, we counted records that were indexed for both gene and disease (double positive hits), for disease only (disease single hits), for gene only (gene single hits), and for neither gene nor disease (double negative hits) to generate a 2 × 2 contingency table. On the basis of the contingency table-framework, we applied different statistical methods to estimate the strength of gene-disease relationships and evaluated the results. These methods included chi-square analysis, Fisher's exact probabilities, relative risk of gene, and relative risk of disease[16] (http://hipseq.med.harvard.edu/MedGene/). In addition, we computed the "product of frequency", which is the product of the proportion of disease/gene double hits to disease single hits and the proportion of disease/gene double hits to gene single hits. To obtain a normal distribution, we transformed all the statistical scores using the natural logarithm. We selected the log of the product of frequency (LPF) to validate MedGene and to use for the analysis with the micro-array data. Spearman rank-correlation coefficients were used to assess the linear relationship between LPF and micro-array fold change in expression level.

Global Analysis. Diseases with at least 50 related genes were selected for clustering analysis, and the LPF scores were normalized with total score for each disease. Hierarchical clustering was done with the "Cluster" software and the clustering result was visualized using "TreeViewer" (http://rana.lbl.gov/EisenSoftware.htm).

Breast Tissue Micro-Arrays. Eighty-nine breast cancer samples (79% ER-positive) and 7 normal breast tissue samples were selected from the Harvard Breast SPORE frozen tissue repository and were representative of the spectrum of histological types, grades, and hormone receptor immuno-phenotypes of breast cancer. Biotinylated cRNA, generated from the total RNA extracted from the bulk tumor, was hybridized to Affymetrix U95A oligo-nucleotide micro-arrays. These micro-arrays consist of 12 400 probes, which represent approximately 9000 genes. Raw expression values were obtained using GENECHIP software from Affymetrix, and then further analyzed using the DNA-Chip Analyzer (dChip) custom software.

Results

Automated Indexing of Medline Records by Disease and Gene. To study the gene-disease associations in the literature, we first compiled complete lists for human diseases and human genes. To index all Medline records that were relevant to human diseases, the Medical Subject Heading (MeSH) index of Medline records was utilized. MeSH is a controlled medical vocabulary from the National Library of Medicine and consists of a set of terms or subject headings that are arranged in both an alphabetic and an hierarchical structure. Medline records are reviewed manually and MeSH terms are added to each with software assistance.[9,10] Twenty-three human disease category headings along with all of their child terms (see the Supporting Information, Supplemental Table 1, or visit http://hipseq.med.harvard.edu/MedGene/publication/s_Table 1.html) were selected from the 2002 MeSH index creating a list of 4033 human diseases.

No index comparable to the MeSH index exists for genes, and thus, it was necessary to apply a string search algorithm for gene names or symbols found in Medline text. A complete list of genes, gene names, gene symbols, and frequently used synonyms were collected from the LocusLink database at NCBI,[11,12] which contains 53 259 independent records keyed by an official gene symbol or name (June 18th, 2002). For the purposes of this study, no distinction was made between genes and their gene products. Authors often use the same name for both, differentiating the two only by the use of italics, if at all. For the intended use of this study, this lack of distinction is unlikely to have a large effect and may in fact be beneficial.

Initial attempts to search the literature using these lists revealed several sources of false positives and false negatives (Table 1). False positives primarily arose when the searched term had other meanings, whereas false negatives arose from syntax discrepancies necessitating the development of filters to reduce these errors. The syntax issues were readily handled by including alternate syntax forms in the search terms. The false positive cases, caused by duplicative and unrelated meanings for the terms, were more difficult to manage. Where possible, case sensitive string mapping reduced inappropriate citations. In many cases, however, this was not sufficient and the terms had to be eliminated entirely, thereby reducing the false positive rate but unavoidably under-representing some genes.

For the purposes of data tracking, a primary gene key was selected to represent all synonyms that correspond to each gene. Medline records were indexed with a primary gene key when any synonym for that key was found in the title or abstract. Case-insensitive string mapping was used for all searches except as noted above. No additional weight was

Table 1. Systematic Sources of False Positives and False Negatives in Unfiltered Data[a]

source of error	error type	example	filter solution
gene symbol/name is not unique	false positive	*MAG*—myelin associated glycoprotein *MAG*—malignancy-associated protein	eliminate this term
gene symbol is unrelated abbreviation	false positive	*PA*—pallid homologue (mouse), pallidin (also abbrev. for Pennsylvania)	eliminate this term
gene symbol/name has language meaning	false positive	*WAS*—Wiskott–Aldrich Syndrome (also the word "was")	case-sensitive string search
nonstandard syntax	false negative	*BAG-1* instead of *BAG1*	add dash term
unofficial gene name/symbol	false negative	*P53* instead of *TP53*	add all gene nicknames
nonspecified gene name	false negative	estrogen receptor instead of Estrogen receptor 1	add family stem term

[a] In preliminary studies, Medline was searched for co-occurrence of genes and diseases and the resulting output was evaluated to identify error sources that were amenable to global filters. Each error source is categorized by the type of error it causes: false positives are suggested relationships that are not real and false negatives are real relationships that are underrepresented. The filter solutions used are indicated. Note that in some cases, the filter solution itself introduces error. In general, error rates maximized sensitivity, even at the expense of specificity if needed.

added for multiple occurrences of a term or the co-occurrence of multiple synonyms for the same gene key.

Medline records were searched with all qualified gene identifiers, such as the official/preferred gene symbol, the official/preferred gene name, all gene nicknames and all syntax variants. In situations where there are several members of a gene family or splice variants, some authors prefer to use a shortened gene family name, e.g., estrogen receptor instead of estrogen receptor 1 (*ESR1*), creating a source of false negatives. For this reason, gene family stem terms were created for all genes that have an alpha or numerical suffix (e.g., *IL2RA*, *TGFβ*, *ESR1*, etc.) and then used to search the literature. The family stem terms were handled separately from the specific gene names so that it would be clear when linkages were made to the gene family versus a specific member in that family.

To improve performance and accuracy, some pre-selection was applied to the records that were scanned. First, review articles were eliminated to avoid redundant treatment of citations. Second, non-English journals were removed because the natural language filters were only relevant to English publications. Finally, journals unlikely to contain primary data about gene-disease relationships were also removed (e.g., *Int. J. Health Educ.*, *Bedside Nurse*, and *J. Health Econ.*). Together, these filters reduced the 12 198 221 Medline publications (July 2002) by 37%.

Ranking the Relative Strengths of Gene-Disease Associations. In total, there were 618 708 gene-disease co-citations, in which 16% (8297) of all studied genes had been associated to a disease and 96% (3875) of all diseases had been associated to at least one gene. To rank the relative strengths of gene disease relationships, we tested several different statistical methods and examined the results. With the exception of the relative risk estimates, the methods provided similar results with respect to the rank order of the gene-disease association strengths. However, after comparing the results to other databases and after consulting disease experts, the log of the product of frequency (LPF) was selected for further analysis because it gave the best results overall.

Validation of MedGene. In developing this tool, it was important to minimize the number of missed genes (false negatives) and miscalled genes (false positives). However, in situations when these goals were in conflict, inclusiveness was prioritized. To determine the false negative rate in MedGene, breast cancer was used as a test case because it was associated with more genes than any other human disease and because

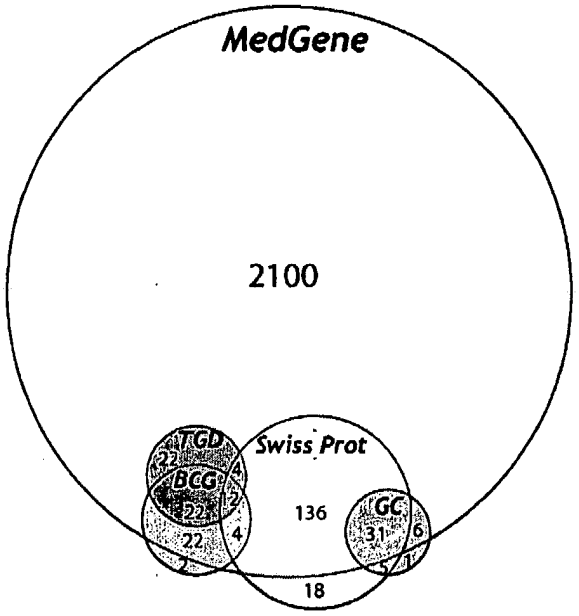


Figure 1. Estimation of the false negative rate by comparison with hand-curated databases. The breast cancer-related genes identified by MedGene were compared with those listed in several other databases including the Tumor Gene Database (TGD),[2] the Breast Cancer Gene Database(BCG),[1] GeneCards (GC)[17] and Swissprot.[18] Genes were considered false negatives if they were represented in at least one of these other databases and not in MedGene and their link to breast cancer was supported by at least one literature reference. All literature references were verified by manual review to confirm their validity. The number of genes in each database or shared by more than one database is indicated. The false negative rate was calculated by genes missed at MedGene (26)/total number of nonoverlapping genes in other databases (285).

there were several public databases that link genes to breast cancer. We compared the list of breast cancer-related genes from MedGene to these databases, illustrated in Figure 1. Among the 285 distinct breast cancer-related genes that were supported by at least one literature citation in these hand-curated databases, 26 were absent from MedGene, suggesting a false negative rate of approximately 9%. To determine why these were missed, all literature references for these genes (80

papers) were reviewed manually (see the Supporting Information, Supplemental Table 2, or visit http://hipseq.med.harvard.edu/MedGene/publication/s_Table 2.html). Among these papers, most false negatives were caused by nonstandard gene terms or gene terms eliminated by our specificity filters. Few genes were missed because they were only mentioned in review papers (0.4%) or they appeared only in the body of the manuscript but not the abstract or title (1.1%). Of note, MedGene identified approximately 2000 additional breast cancer-related genes not listed in any other database.

To assess the false positive error rate, two complementary approaches were used: a detailed analysis of one disease and a global examination of 1000 diseases. The detailed approach examined the false positive error rate and its sources, whereas the global approach tested whether the overall results made biomedical sense.

Using the LPF, 1467 genes related to prostate cancer were assembled in rank order. We then retrieved approximately 300 Medline records each for the highest ranked 100 and the lowest ranked 200 genes and manually reviewed the titles and abstracts to determine the verity of the association. Nearly 80% of the highest ranked 100 genes fell into one of the five categories that reflect meaningful gene-disease relationships (see the Supporting Information, Supplemental Table 3, or visit http://hipseq.med.harvard.edu/MedGene/publication/s_Table 3.html). Among the lowest ranked 200 genes, approximately 70% reflected true relationships. Of the 600 records reviewed, there were only two in which the association between the gene and the disease was described as negative. Both were genes with very low scores. In both cases, the authors did not argue the absence of *any* relationship, but rather that a particular feature of the gene or protein was not shown to be related to human prostate cancer.[13,14]

The coincidence of some gene symbols with medical abbreviations, chemical abbreviations and biological abbreviations resulted in most of the false positives (see the Supporting Information, Supplemental Table 4, or visit http://hipseq.med.harvard.edu/MedGene/publication/s_Table 4.html), emphasizing the importance of the filters that were added in the search algorithm (Table 1). Without the filters, the false positive rate more than doubled, and the false negative rate rose dramatically (data not shown). For example, among the papers about breast cancer, there were only 12 Medline records that referred to *ESR1* and 10 to *ESR2*, whereas almost 2000 papers mentioned estrogen receptor without specifying *ESR1* or *ESR2*; this latter group was detected by the family stem term filter.

To further validate these results, a global analysis of the gene-disease relationships described by MedGene was performed. For this experiment, it was reasoned that the more closely related the diseases are to one another, the more they will be related to the same gene sets. Thus, if the relationships defined by MedGene accurately reflected the literature, then an unsupervised hierarchical clustering of the gene data should group diseases in a manner consistent with common medical thinking. Conversely, if the clustered diseases do not make sense biologically or medically, it may reflect excessive false positives, false negatives, or inappropriate scoring of the data.

To execute this experiment, the gene sets and the corresponding LPF values for 1000 randomly selected diseases (each with at least 50 gene relationships) were used as a dataset for clustering the diseases. A review of the results showed that the resulting disease clusters were indeed logical based upon common medical knowledge (see the Supporting Information, Supplemental Figure 1, or visit http://hipseq.med.harvard.edu/MedGene/publication/s_Figure 1.html). For example, in one such cluster shown in Figure 2, diabetes and its complications grouped together and were also closely linked to diseases associated with starvation states.

The number of genes associated with a given disease can be estimated by adjusting the MedGene number up by the false negative rate (~9%) and down by the false positive rate (~26% on average). Using this, the average disease has 103.7 ± 45.3 (mean ± s.d.) genes associated with it, although the range is quite broad with 2359 genes related to breast cancer, 2122 genes related to lung cancer and no genes related to a number of diseases.

Applying MedGene to the Analysis of Large Datasets. Access to a comprehensive summary of the genes linked to human diseases provided an opportunity to analyze data obtained from a high-throughput experiment. We compared the MedGene breast cancer gene list to a gene expression data set generated from a micro-array analysis comparing breast cancer and normal breast tissue samples. Micro-array analysis identified 2286 genes that had greater than a 1-fold difference in mean expression level between breast cancer samples and normal breast samples. Using MedGene, we sorted the 2286 genes into four classes: 555 genes directly linked to breast cancer in the literature by gene term search (first-degree association by gene name); 328 genes directly linked by family term search (first-degree association by family term); 1021 genes linked to breast cancer only through other breast cancer genes (second-degree association); and 505 genes not previously associated with breast cancer. (See the Supporting Information, Supplemental Figure 2, or visit http://hipseq.med.harvard.edu/MedGene/publication/s_Figure 2.html.) Among the 505 previously unrelated genes, 467 were either newly identified genes or genes that had not previously been associated with any disease. Among the remaining 38 genes, 9 had been related to other cancers, specifically esophageal, colon, uterine, skin, and cervix.

To determine whether the genes highlighted by the micro-array analysis were more likely to have been previously linked to breast cancer in the literature, we created a two-dimensional plot of the fold change of expression level between breast cancer and normal tissue versus the literature score (LPF) (Figure 3A). There was a broad spread of expression changes among the genes directly linked to breast cancer ranging from less than 1-fold change (68%) to over 40-fold (0.3%). Notably, the majority of genes with greater than 10-fold expression changes were linked to breast cancer by first-degree association.

Among all 754 genes directly linked to breast cancer in the literature, there was no correlation between LPF and micro-array fold change ($r = 0.018$, p-value = 0.62). However, when we stratified the analysis based on the magnitude of the fold change, we observed an increasing trend in correlation (Figure 3B) suggesting that genes with a more substantial change in expression level were more likely to have a stronger association in the literature. For genes that had 10-fold change or more in expression level, the correlation increased to 0.41 (p-value = 0.05).

When we evaluated the micro-array data separately for ER positive and ER negative tumors, the trend in correlation between fold change and literature score was highly dependent on estrogen receptor status. Interestingly, there was a similar trend in correlation for ER positive tumors, but no trend in correlation for ER negative tumors.

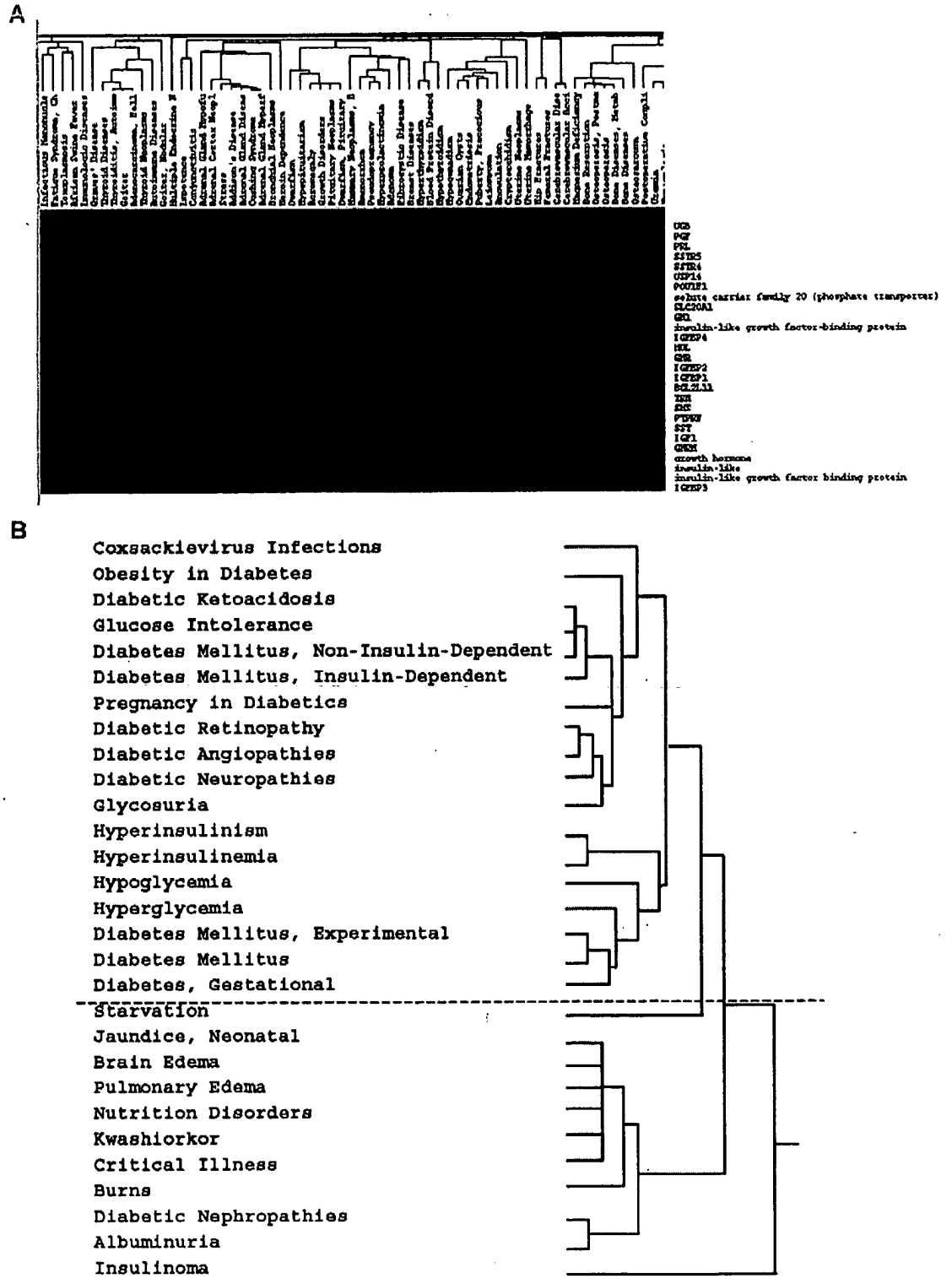


Figure 2. Global validation by clustering analysis. 2(**A**). The gene sets and the corresponding LPF values for 1000 diseases, each with at least 50 gene relationships, were used in an unsupervised clustering of the diseases based on the gene patterns associated with them. A sample of the data is shown here. 2(**B**). One of the resulting clusters is shown that corresponds to blood sugar states. Diabetes terms (above the line) and starvation states terms (under the line) clustered together. Within these groups, there is also clustering of diabetic small vessel complications, altered serum chemistries, nutritional disorders, etc.(Supplemental Figure 1: http://hipseq.med.harvard.edu/MedGene/publication/s_Figure 1.html).

Finally, to validate our findings, we computed similar correlations between the breast cancer expression data and LPF scores generated by MedGene for hypertension, a disease unrelated to breast cancer. As expected, we did not observe an increasing trend in correlation for hypertension.

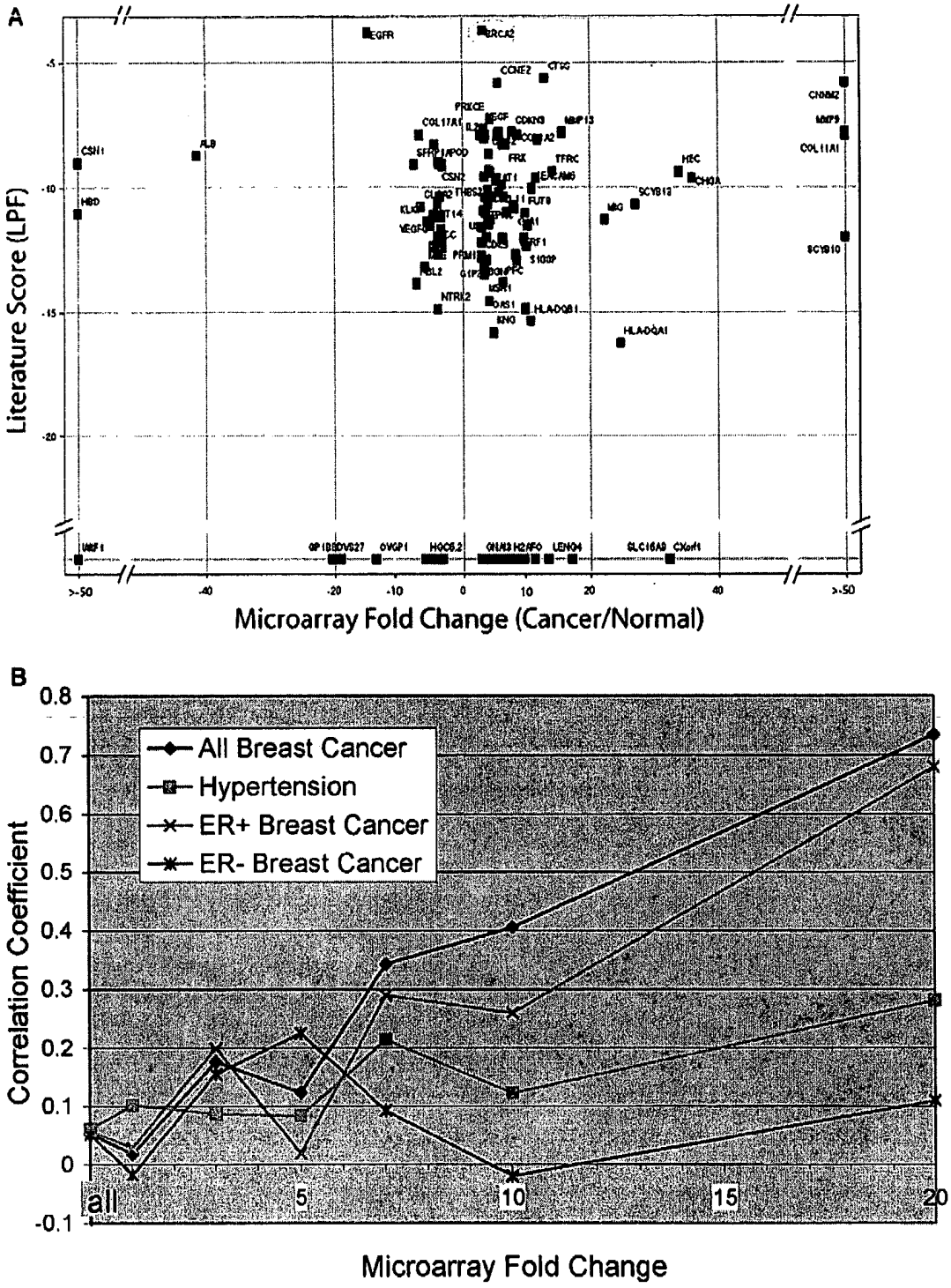


Figure 3. Relationship between literature score and functional data for breast cancer. **3A.** The data from an expression analysis of samples for breast tumors and normal breast tissue were analyzed to indicate the fold difference of expression level between breast tumor and normal sample (cutoff ≥ 3-fold change). The fold changes were plotted against the literature score for the same gene set. Green dots represent first-degree association by gene search, blue dots represent first-degree association by family search and red dots represent no-association. Some well-studied genes, such as BRCA2 (pink circle), are not reflected by a substantial difference in expression level. Furthermore, the majority of genes that have no association with breast cancer in the literature had less than 10-fold expression changes (shaded area). **3B.** The Spearman rank-correlation coefficients between literature score (LPF) and the fold change of expression level between tumor and normal breast samples (*y*-axis) in relation to the amount of fold change of expression level (*x*-axis). Gene rank lists were generated for breast cancer (blue) and hypertension (pink). Correlations were also computed between the breast cancer gene LPF scores and fold change expression data among estrogen receptor positive tumors only (light blue) and estrogen receptor negative tumors only (purple).

Table 2. Top 25 Genes Related to Selected Human Diseases[a]

breast neoplasms	hypertension	rheumatoid arthritis	bipolar disorder	atherosclerosis
estrogen receptor	*REN*	*RA*	*ERDA1*	apolipoprotein
PGR	*DBP*	*TNFRSF10A*	*SNAP29*	*APOE*
ERBB2	*LEP*	*CRP*	*PFKL*	*LDLR*
BRCA1	*AGT*	*AS*	*DRD2*	*ELN*
BRCA2	*INS*	*ESR1*	*TRH*	*ARG1*
EGFR	kallikrein	*HLA-DRB1*	*IMPA2*	*APOB*
CYP19	*ACE*	*DR1*	*HTR3A*	*APOA1*
TFF1	endothelin	interleukin	*DRD3*	*MSR1*
PSEN2	*S100A6*	*TNF*	*REM*	*LPL*
TP53	*BDK*	*IL6*	*KCNN3*	*PON1*
CES3	*DIANPH*	collagen	*DRD4*	plasminogen activator inhibitor
CEACAM5	*SAR1*	*IL1A*	*HTR2C*	*PLG*
ERBB3	*PIH*	*ACR*	*RELN*	vascular cell adhesion molecule
cyclin	*CD59*	*TNFRSF12*	*DBH*	*ATOH1*
COX5A	*ALB*	*IL2*	*MAOA*	*VWF*
cathepsin	*CYP11B2*	*CHI3L1*	*COMT*	*INS*
ERBB4	*MAT2B*	*IL8*	*HTR2A*	*ARG2*
TRAM	angiotensin receptor	interleukin 1	*SYNJ1*	*ABCA1*
CCND1	*AGTR2*	matrix metalloproteinase	*INPP1*	*OLR1*
EGF	*NPPA*	interferon	*NEDD4L*	collagen
MUC1	*LVM*	*CD68*	*FRA13C*	*MCP*
insulin-like	*DBH*	*IL4*	transducer of *ERBB2*	lipoprotein
BCL2	*NPY*	*IL17*	*BAIAP3*	*APOA2*
mucin	*POMC*	*MMP3*	*ATP1B3*	intercellular adhesion molecule
FGF3	neuropeptide	*SIL*	*DRD5*	*RAB27A*

[a] MedGene results for the top 25 genes associated with breast neoplasms, hypertension, rheumatoid arthritis, bipolar disorder, and atherosclerosis, respectively, ranked by LPF scores. The hyperlink to all the papers co-citing the gene and the disease is available at MedGene website (http://hipseq.med.harvard.edu/MedGene/).

Discussion

The Human Genome Project heralded a new era in biological research where the emphasis on understanding specific pathways has expanded to global studies of genomic organization and biological systems. High-throughput technologies can provide novel insight into comprehensive biological function but also introduces new challenges. The utility of these technologies is limited to the ability to generate, analyze, and interpret large gene lists. MedGene, a relational database derived by mining the information in Medline, was created to address this need. MedGene users can query for a rank-ordered list of human gene-disease relationships (Table 2) for one or more diseases. Each entry is hyperlinked to the original papers supporting each association and to other relevant databases.

MedGene is an innovative extension of previous text mining approaches. Perez-Iratxeta et al. used the GO annotation and their chromosomal locations to predict genes that may contribute to inherited disorders.[8] MedGene takes a broader view and includes all diseases and all possible gene-disease relationships. Furthermore, MedGene utilizes co-citation to indicate a relationship rather than GO annotation, which is limited to the subset of genes that have GO annotation. Our approach is complementary to that taken by Chaussabel and Sher, who used the frequency of co-cited terms to cluster genes into a hierarchy of gene-gene relationships.[6]

A unique aspect of this tool is the ability to assess the relative strengths of gene-disease relationships based on the frequency of both co-citation and single citation. This presupposes that most co-citations describe a positive association, often referred to as publication bias[15] and is supported by our observations that negative associations are rare (Supplemental Table 3: http://hipseq.med.harvard.edu/MedGene/publication/s_Table 3.html). Of course, relationships established by frequency of co-citation do not necessarily represent a true biological link; however, it is strong evidence to support a true relationship.

Another important feature of MedGene is the implementation of software filters that substantially reduced the error rate. We estimate that less than 10% of all associations were missed and at least 70% of even the weakest associations were real. For this study, all of the filters that we applied were general ones, e.g., expanding the list of all gene names to address the different syntax forms used by different journals, eliminating gene names that correspond to common English words, etc. The majority of the remaining search term ambiguities were idiosyncratic and difficult to identify systematically without causing a significant rise in false negatives. Alternative approaches, such as the examination of the nearest neighbor terms, need to be considered to further reduce the false positive rate.

It is not uncommon to see expression changes in microarray experiments as small as 2-fold reported in the literature. Even when these expression changes are statistically significant, it is not always clear if they are biologically meaningful. When comparing expression levels of disease to normal tissue, one expects an enrichment of known disease-related genes to appear in the altered expression group. MedGene provided a unique opportunity to test this notion in the context of existing knowledge on a novel breast cancer micro-array dataset. For genes displaying a 5-fold change or less in tumors compared to normal, there was no evidence of a correlation between altered gene expression and a known role in the disease. This

Table 3. Genes with Large Expression Changes in ER− but Not in ER+ Breast Tumors

gene symbol	fold change (ER+)	fold change (ER−)
KRTHB1	1.0	610.8
BRS3	1.2	89.4
DKK1	1.2	69.8
ZIC1	1.9	59.6
TLR1	1.0	38.5
KIAA0680	2.6	33.2
CDKN3	1.0	30.6
EBI2	4.0	27.9
GZMB	3.8	21.9
STK18	4.7	18.6
GPR49	1.0	14.6
MYO10	1.6	14.4
LAD1	−1.0	13.5
POLE2	4.2	13.0
HMG4	4.4	12.9
BCL2L11	−1.2	12.3
LRP8	2.9	12.2
CCNB2	1.0	11.8
CCNE2	4.0	11.6
FGB	−4.3	11.1
KNSL6	2.9	10.9
H1F5	3.0	10.2
SERPINH2	4.6	10.2
YAP1	1.0	10.0
LPHB	−1.3	−10.4
TCEA2	−1.1	−10.8
TFF1	1.3	−11.4
COL17A1	−4.1	−15.7
POP5	1.1	−16.2
BPAG1	−4.6	−22.3
PDZK1	−1.1	−36.8
VEGFC	−2.8	−51.5
MUC6	−1.4	−64.9
SERPINA5	−1.0	−83.1
MEIS1	−1.6	−85.9
CA12	2.4	−150.3

Table 3. MedGene identified a set of relatively understudied, yet highly expressed genes in ER negative, but not ER positive breast tumors. All of these genes have either never been co-cited with breast cancer or have a weak association except those marked with an *.

reflects the many genes whose role in breast cancer may not involve large changes in expression in sporadic tumors (e.g., *BRCA1* and *BRCA2*) and genes whose modest changes in expression may be unrelated to the disease. Strikingly, among genes with a 10-fold change or more in expression level, there was a strong and significant correlation between expression level and a published role in the disease, providing the first global validation of the micro-array approach to identifying disease-specific genes.

The results derived from MedGene have two implications. First, a careful hunt for corroborating evidence of a role in breast cancer should precede any further study of genes with less than 5-fold expression level changes. Second, any genes with 10-fold changes or more are likely to be related to breast cancer and warrant attention. It is likely that this threshold will change depending on the disease as well as the experiment.

Interestingly, the observed correlation was only found among ER-positive tumors, not ER-negative. This may reflect a bias in the literature to study the more prevalent type of tumor in the population. Furthermore, this emphasizes that caution must be taken when interpreting experiments that may contain subpopulations that behave very differently. The MedGene approach identified a set of relatively understudied, yet highly expressed genes in ER-negative tumors that are worthy of further examination (Table 3).

In conclusion, we have developed an automated method of summarizing and organizing the vast biomedical literature. To our knowledge, the resulting database is the most comprehensive and accurate of its kind. By generating a score that reflects the strength of the association, it provides an important tool for the rapid and flexible analysis of large datasets from various high-throughput screening experiments. Furthermore, it can be used for selecting subsets of genes for functional studies, for building disease-specific arrays, for looking at genes common to multiple diseases and various other high-throughput applications. In the future, it will be possible to enhance the utility of the MedGene database by building links between genes and other MeSH terms as well as other biological processes and concepts, such as cell division and responses to small molecules.

Acknowledgment. We would like to thank P. Braun, L. Garraway, J. Pearlberg, and other members of our institute for helpful discussion. Many thanks to the NLM (National Library of Medicine) for licensing of MEDLINE and the annotation effort of adding MeSH indexes for MEDLINE abstracts. This work was funded by grants from the Breast Cancer Research Foundation and an NHLBI PGA Grant (Vol HL66582-02).

Supporting Information Available: Twenty-three human disease category headings along with all of their child terms selected from the 2002 MeSH index (Supplemental Table 1); analysis of the causes of false negatives in MedGene (Supplemental Table); meaningful gene-disease relationships found in MedGene (Supplemental Table 3); causes for incorrect assignment of gene indexes (Supplemental Table 4); a review of the results, showing that the resulting disease clusters were indeed logical (Supplemental Figure 1); and a review of the results showing that among the 505 previously unrelated genes, 467 were either newly identified genes or genes that had not previously been associated with any disease (Supplemental Figure 2). This material is available free of charge via the Internet at http://pubs.acs.org and at the web sites mentioned in the text.

References

(1) Baasiri, R. A.; Glasser, S. R.; Steffen, D. L.; Wheeler, D. A. *Oncogene* **1999**, *18*, 7958−7965.
(2) Steffen, D. L.; Levine, A. E.; Yarus, S.; Baasiri, R. A.; Wheeler, D. A. *Bioinformatics* **2000**, *16*, 639−649.
(3) Marcotte, E. M.; Xenarios, I.; Eisenberg, D. *Bioinformatics* **2001**, *17*, 359−363.
(4) Ono, T.; Hishigaki, H.; Tanigami, A.; Takagi, T. *Bioinformatics* **2001**, *17*, 155−161.
(5) Jenssen, T. K.; Laegreid, A.; Komorowski, J.; Hovig, E. *Nat. Genet.* **2001**, *28*, 21−28.
(6) Chaussabel, D.; Sher, A. *Genome Biol.* **2002**, *3*, RESEARCH0055.
(7) Gibbons, F. D.; Roth, F. P. *Genome Res.* **2002**, *12*, 1574−1581.
(8) Perez-Iratxeta, C.; Bork, P.; Andrade, M. A. *Nat. Genet.* **2002**, *31*, 316−319.
(9) Funk, M. E.; Reid, C. A. *Bull. Med. Libr. Assoc.* **1983**, *71*, 176−183.
(10) Humphrey, S. M.; Miller, N. E. *J. Am. Soc. Inf. Sci.* **1987**, *38*, 184−196.
(11) Maglott, D. R.; Katz, K. S.; Sicotte, H.; Pruitt, K. D. *Nucleic Acids Res.* **2000**, *28*, 126−128.
(12) Pruitt, K. D.; Maglott, D. R. *Nucleic Acids Res.* **2001**, *29*, 137−140.
(13) Wadelius, M.; Andersson, A. O.; Johansson, J. E.; Wadelius, C.; Rane, E. *Pharmacogenetics* **1999**, *9*, 333−340.
(14) Adam, R. M.; Borer, J. G.; Williams, J.; Eastham, J. A.; Loughlin, K. R.; Freeman, M. R. *Endocrinology* **1999**, *140*, 5866−5875.
(15) Montori, V. M.; Smieja, M.; Guyatt, G. H. *Mayo Clin. Proc.* **2000**, *75*, 1284−1288.
(16) Denenberg, V. H. *Statistics Experimental Design for Behavioral and Biological Researchers*; Wiley-Liss: New York, 1976.
(17) Rebhan, M.; Chalifa-Caspi, V.; Prilusky, J.; Lancet, D. *Trends Genet.* **1997**, *13*, 163.
(18) Bairoch, A.; Apweiler, R. *Nucleic Acids Res.* **2000**, *28*, 45−48.

PR0340227

Review

Paul A. Haynes
Steven P. Gygi
Daniel Figeys
Ruedi Aebersold

Department of Molecular Biotechnology, University of Washington, Seattle, WA, USA

Proteome analysis: Biological assay or data archive?

In this review we examine the current state of proteome analysis. There ar three main issues discussed: why it is necessary to study proteomes; how pro teomes can be analyzed with current technology; and how proteome analysi can be used to enhance biological research. We conclude that proteome anal ysis is an essential tool in the understanding of regulated biological systems Current technology, while still mostly limited to the more abundant proteins enables the use of proteome analysis both to establish databases of protein present, and to perform biological assays involving measurement of multipl variables. We believe that the utility of proteome analysis in future biologica research will continue to be enhanced by further improvements in analytica technology.

Contents

1 Introduction 1862
2 Rationale for proteome analysis 1862
2.1 Correlation between mRNA and protein expression levels 1863
2.2 Proteins are dynamically modified and processed 1863
2.3 Proteomes are dynamic and reflect the state of a biological system 1863
3 Description and assessment of current proteome analysis technology 1863
3.1 Technical requirements of proteome technology 1863
3.2 2D electrophoresis – mass spectrometry: a common implementation of proteome analysis 1864
3.3 Protein identification by LC-MS/MS, capillary LC-MS/MS and CE-MS/MS 1865
3.3.1 LC-MS/MS 1865
3.3.2 Capillary LC-MS 1865
3.3.3 CE-MS/MS 1865
3.4 Assessment of 2-DE-MS proteome technology 1866
4 Utility of proteome analysis for biological research 1868
4.1 The proteome as a database 1868
4.2 The proteome as a biological assay 1868
5 Concluding remarks 1870
6 References 1870

1 Introduction

A proteome has been defined as the protein complement expressed by the genome of an organism, or, in multicellular organisms, as the protein complement expressed by a tissue or differentiated cell [1]. In the most common implementation of proteome analysis the proteins extracted from the cell or tissue analyzed are separated by high resolution two-dimensional gel electrophoresis (2-DE) detected in the gel and identified by their amino acid sequence. The ease, sensitivity and speed with which gel separated proteins can be identified by the use of recently developed mass spectrometric techniques have dramatically increased the interest in proteome technology. One of the most attractive features of such analyses is that complex biological systems can potentially be studied in their entirety, rather than as a multitude of individual components. This makes it far easier to uncover the many complex, and often obscure, relationships between mature gene products in cells. Large-scale proteome characterization projects have been undertaken for a number of different organisms and cell types. Microbial proteome projects currently in progress include, for example: *Saccharomyces cerevisiae* [2], *Salmonella enterica* [3], *Spiroplasma melliferum* [4], *Mycobacterium tuberculosis* [5], *Ochrobactrum anthropi* [6], *Haemophilus influenzae* [7], *Synechocystis spp.* [8], *Escherichia coli* [9], *Rhizobium leguminosarum* [10], and *Dictyostelium discoideum* [11]. Proteome projects underway for tissues of more complex organisms include those for: human bladder squamous cell carcinomas [12], human liver [13], human plasma [13], human keratinocytes [12], human fibroblasts [12], mouse kidney [12], and rat serum [14]. In this manuscript we critically assess the concept of proteome analysis and the technical feasibility of establishing complete proteome maps, and discuss ways in which proteome analysis and biological research intersect.

2 Rationale for proteome analysis

The dramatic growth in both the number of genome projects and the speed with which genome sequences are being determined has generated huge amounts of sequence information, for some species even complete genomic sequences ([15–17]). The description of the state of a biological system by the quantitative measurement of system components has long been a primary objective in molecular biology. With recent technical advances including the development of differential display-PCR [18], cDNA microarray and DNA chip technology [19, 20] and serial analysis of gene expression (SAGE) [21, 22], it is now feasible to establish global and quantitative mRNA expression maps of cells and tissues in which the sequence of all the genes is known, at a speed and sensitivity which is not matched by current

Correspondence: Professor Ruedi Aebersold, Department of Molecular Biotechnology, University of Washington, Box 357730, Seattle, WA, 98195, USA (Tel: +206-685-4235; Fax: +206-685-6392; E-mail: ruedi@u.washington.edu)

Abbreviations: CID, collision-induced dissociation; MS/MS, tandem mass spectrometry; SAGE, serial analysis of gene expression

Keywords: Proteome / Two-dimensional polyacrylamide gel electrophoresis / Tandem mass spectrometry

© WILEY-VCH Verlag GmbH, 69451 Weinheim, 1998

0173-0835/98/1111-1862 $17.50+.50/0

protein analysis technology. Given the long-standing paradigm in biology that DNA synthesizes RNA which synthesizes protein, and the ability to rapidly establish global, quantitative mRNA expression maps, the questions which arise are why technically complex proteome projects should be undertaken and what specific types of information could be expected from proteome projects which cannot be obtained from genomic and transcript profiling projects. We see three main reasons for proteome analysis to become an essential component in the comprehensive analysis of biological systems. (i) Protein expression levels are not predictable from the mRNA expression levels, (ii) proteins are dynamically modified and processed in ways which are not necessarily apparent from the gene sequence, and (iii) proteomes are dynamic and reflect the state of a biological system.

2.1 Correlation between mRNA and protein expression levels

Interpretations of quantitative mRNA expression profiles frequently implicitly or explicitly assume that for specific genes the transcript levels are indicative of the levels of protein expression. As part of an ongoing study in our laboratory, we have determined the correlation of expression at the mRNA and protein levels for a population of selected genes in the yeast *Saccharomyces cerevisiae* growing at mid-log phase (S. P. Gygi *et al.*, submitted for publication). mRNA expression levels were calculated from published SAGE frequency tables [22]. Protein expression levels were quantified by metabolic radiolabeling of the yeast proteins, liquid scintillation counting of the protein spots separated by high resolution 2-DE and mass spectrometric identification of the protein(s) migrating to each spot. The selected 80 samples constitute a relatively homogeneous group with respect to predicted half-life and expression level of the protein products. Thus far, we have found a general trend but no strong correlation between protein and transcript levels (Fig. 1). For some genes studied equivalent mRNA transcript levels translated into protein abundances which varied by more than 50-fold. Similarly, equivalent steady-state protein expression levels were maintained by transcript levels varying by as much as 40-fold (S. P. Gygi *et al.*, submitted). These results suggests that even for a population of genes predicted to be relatively homogeneous with respect to protein half-life and gene expression, the protein levels cannot be accurately predicted from the level of the corresponding mRNA transcript.

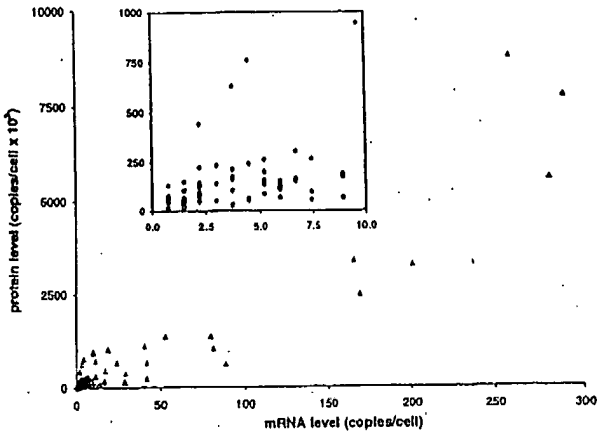


Figure 1. Correlation between mRNA and protein levels in yeast cells. For a selected population of 80 genes, protein levels were measured by ^{35}S-radiolabeling and mRNA levels were calculated from published SAGE tables. Inset: expanded view of the low abundance region. For more experimental details, also see Figs. 5 and 6, (S. P. Gygi *et al.*, submitted).

2.2 Proteins are dynamically modified and processed

In the mature, biologically active form many proteins are post-translationally modified by glycosylation, phosphorylation, prenylation, acylation, ubiquitination or one or more of many other modifications [23] and many proteins are only functional if specifically associated or complexed with other molecules, including DNA, RNA, proteins and organic and inorganic cofactors. Frequently, modifications are dynamic and reversible and may alter the precise three-dimensional structure and the state of activity of a protein. Collectively, the state of modification of the proteins which constitute a biological system are important indicators for the state of the system. The type of protein modification and the sites modified at a specific cellular state can usually not be determined from the gene sequence alone.

2.3 Proteomes are dynamic and reflect the state of a biological system

A single genome can give rise to many qualitatively and quantitatively different proteomes. Specific stages of the cell cycle and states of differentiation, responses to growth and nutrient conditions, temperature and stress, and pathological conditions represent cellular states which are characterized by significantly different proteomes. The proteome, in principle, also reflects events that are under translational and post-translational control. It is therefore expected that proteomics will be able to provide the most precise and detailed molecular description of the state of a cell or tissue, provided that the external conditions defining the state are carefully determined. In answer to the question of whether the study of proteomes is necessary for the analysis of biomolecular systems, it is evident that the analysis of mature protein products in cells is essential as there are numerous levels of control of protein synthesis, degradation, processing and modification, which are only apparent by direct protein analysis.

3 Description and assessment of current proteome analysis technology

3.1 Technical requirements of proteome technology

In biological systems the level of expression as well as the states of modification, processing and macro-molecular association of proteins are controlled and modulated depending on the state of the system. Comprehensive analysis of the identity, quantity and state of modification of proteins therefore requires the detection and

quantitation of the proteins which constitute the system, and analysis of differentially processed forms. There are a number of inherent difficulties in protein analysis which complicate these tasks. First, proteins cannot be amplified. It is possible to produce large amounts of a particular protein by over-expression in specific cell systems. However, since many proteins are dynamically post-translationally modified, they cannot be easily amplified in the form in which they finally function in the biological system. It is frequently difficult to purify from the native source sufficient amounts of a protein for analysis. From a technological point of view this translates into the need for high sensitivity analytical techniques. Second, many proteins are modified and processed post-translationally. Therefore, in addition to the protein identity, the structural basis for differentially modified isoforms also needs to be determined. The distribution of a constant amount of protein over several differentially modified isoforms further reduces the amount of each species available for analysis. The complexity and dynamics of post-translational protein editing thus significantly complicates proteome studies. Third, proteins vary dramatically with respect to their solubility in commonly used solvents. There are few, if any, solvent conditions in which all proteins are soluble and which are also compatible with protein analysis. This makes the development of protein purification methods particularly difficult since both protein purification and solubility have to be achieved under the same conditions. Detergents, in particular sodium dodecyl sulfate (SDS), are frequently added to aqueous solvents to maintain protein solubility. The compatibility with SDS is a big advantage of SDS polyacrylamide gel electrophoresis (SDS-PAGE) over other protein separation techniques. Thus, SDS-PAGE and two-dimensional gel electrophoresis, which also uses SDS and other detergents, are the most general and preferred methods for the purification of small amounts of proteins, provided that activity does not necessarily need to be maintained. Lastly, the number of proteins in a given cell system is typically in the thousands. Any attempt to identify and categorize all of these must use methods which are as rapid as possible to allow completion of the project within a reasonable time frame. Therefore, a successful, general proteomics technology requires high sensitivity, high throughput, the ability to differentiate differentially modified proteins, and the ability to quantitatively display and analyze all the proteins present in a sample.

3.2 2-D electrophoresis – mass spectrometry: a common implementation of proteome analysis

The most common currently used implementation of proteome analysis technology is based on the separation of proteins by two-dimensional (IEF/SDS-PAGE) gel electrophoresis and their subsequent identification and analysis by mass spectrometry (MS) or tandem mass spectrometry (MS/MS). In 2-DE, proteins are first separated by isoelectric focusing (IEF) and then by SDS-PAGE, in the second, perpendicular dimension. Separated proteins are visualized at high sensitivity by staining or autoradiography, producing two-dimensional arrays of proteins. 2-DE gels are, at present, the most commonly used means of global display of proteins in complex samples. The separation of thousands of proteins has been achieved in a single gel [24, 25] and differentially modified proteins are frequently separated. Due to the compatibility of 2-DE with high concentrations of detergents, protein denaturants and other additives promoting protein solubility, the technique is widely used.

The second step of this type of proteome analysis is the identification and analysis of separated proteins. Individual proteins from polyacrylamide gels have traditionally been identified using *N*-terminal sequencing [26, 27], internal peptide sequencing [28, 29], immunoblotting or comigration with known proteins [30]. The recent dramatic growth of large-scale genomic and expressed sequence tag (EST) sequence databases has resulted in a fundamental change in the way proteins are identified by their amino acid sequence. Rather than by the traditional methods described above, protein sequences are now frequently determined by correlating mass spectral or tandem mass spectral data of peptides derived from proteins, with the information contained in sequence databases [31–33].

There are a number of alternative approaches to proteome analysis currently under development. There is considerable interest in developing a proteome analysis stragegy which bypasses 2-DE altogether, because it is considered a relatively slow and tedious process, and because of perceived difficulties in extracting proteins from the gel matrix for analysis. However, 2-DE as a starting point for proteome analysis has many advantages compared to other techniques available today. The most significant strengths of the 2-DE-MS approach include the relatively uniform behavior of proteins in gels, the ability to quantify spots and the high resolution and simultaneous display of hundreds to thousands of proteins within a reasonable time frame.

A schematic diagram of a typical procedure of the identification of gel-separated proteins is shown in Fig. 2. Protein spots detected in the gel are enzymatically or chemically fragmented and the peptide fragments are isolated for analysis, as already indicated, most frequently by MS or MS/MS. There are numerous protocols for the generation of peptide fragments from gel-separated proteins. They can be grouped into two categories, digestion in the gel slice [28, 34] or digestion after electrotransfer out of the gel onto a suitable membrane ([29, 35–37] and reviewed in [38]). In most instances either technique is applicable and yields good results. The analysis of MS or MS/MS data is an important step in the whole process because MS instruments can generate an enormous amount of information which cannot easily be managed manually. Recently, a number of groups have developed software systems dedicated to the use of peptide MS and MS/MS spectra for the identification of proteins. Proteins are identified by correlating the information contained in the MS spectra of protein digests or MS/MS spectra of individual peptides with data contained in DNA or protein sequence databases.

The systems we are currently using in our laboratory are based on the separation of the peptides contained in protein digests by narrow bore or capillary liquid chromatog-

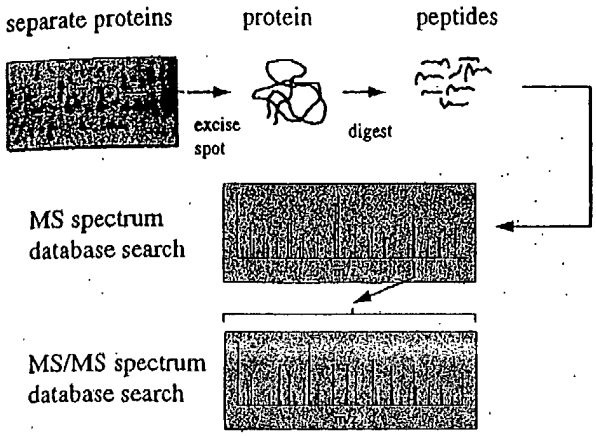


Figure 2. Schematic diagram of a procedure for identification of gel-separated proteins. Peptides can either be separated by a technique such as LC or CE, or infused as a mixture and sorted in the MS. Database searching can either be performed on peptide masses from an MS spectrum, peptide fragment masses from CID spectra of peptides, or a combination of both.

raphy [39, 40] or capillary electrophoresis [41], the analysis of the separated peptides by electrospray ionization (ESI) MS/MS, and the correlation of the generated peptide spectra with sequence databases using the SEQUEST program developed at the University of Washington [32, 33]. The system automatically performs the following operations: a particular peptide ion characterized by its mass-to-charge ratio is selected in the MS out of all the peptide ions present in the system at a particular time; the selected peptide ion is collided in a collision cell with argon (collision-induced dissociation, CID) and the masses of the resulting fragment ions are determined in the second sector of the tandem MS; this experimentally determined CID spectrum is then correlated with the CID spectra predicted from all the peptides in a sequence database which have essentially the same mass as the peptide selected for CID; this correlation matches the isolated peptide with a sequence segment in a database and thus identifies the protein from which the peptide was derived. There are a number of alternative programs which use peptide CID spectra for protein identification, but we use the SEQUEST system because it is currently the most highly automated program and has proven to be successful, versatile and robust.

3.3 Protein identification by LC-MS/MS, capillary LC-MS/MS and CE-MS/MS

It has been demonstrated repeatedly that MS has a very high intrinsic sensitivity. For the routine analysis of gel-separated proteins at high sensitivity, the most significant challenge is the handling of small amounts of sample. The crux of the problem is the extraction and transferal of peptide mixtures generated by the digestion of low nanogram amounts of protein, from gels into the MS/MS system without significant loss of sample or introduction of unwanted contaminants. We employ three different systems for introducing gel-purified samples into an MS, depending on the level of sensitivity required. As an approximate guideline, for samples containing tens of picomoles of peptides, LC-MS/MS is most appropriate; for samples containing low picomole amounts to high femtomole amounts we use capillary LC-MS/MS; and for samples containing femtomoles or less, CE-MS/MS is the method of choice.

3.3.1 LC-MS/MS

The coupling of an MS to an HPLC system using a 0.5 mm diameter or bigger reverse phase (RP) column has been described in detail [42]. This system has several advantages if a large number of samples are to be analyzed and all are available in sufficient quantity. The LC-MS and database searching program can be run in a fully automated mode using an autosampler, thus maximizing sample throughput and minimizing the need for operator interference. The relatively large column is tolerant of high levels of impurities from either gel preparation or sample matrix. Lastly, if configured with a flow-splitter and micro-sprayer [40], analyses can be performed on a small fraction of the sample (less than 5%) while the remainder of the sample is recovered in very pure solvents. This latter feature is particularly useful when an orthogonal technique is also used to analyze peptide fractions, such as scintillation of an introduced radiolabel, and this data can be correlated with peptides identified by CID spectra.

3.3.2 Capillary LC-MS

An increase of sensitivity of approximately tenfold can be achieved by using a capillary LC system with a 100 μm ID column rather than a 0.5 mm ID column as referred to above. Since very low flow rates are required for such columns, most reports have used a precolumn flow splitting system for producing solvent gradients. We have recently desribed the design and construction of a novel gradient mixing system which enables the formation of reproducible gradients at very low flow rates (low nL/min) without the need for flow splitting (A. Ducret *et al.*, submitted for publication). Using this capillary LC-MS/MS system we were able to identify gel-separated proteins if low picomole to high femtomole amounts were loaded onto the gel [40]. This system is as yet not automated and, like all capillary LC systems, is prone to blockage of the columns by microparticulates when analyzing gel-separated proteins.

3.3.3 CE-MS/MS

The highest level of sensitivity for analyzing gel-separated proteins can be achieved by using capillary electrophoresis – mass spectrometry (CE-MS). We have described in the past a solid-phase extraction capillary electrophoresis (SPE-CE) system which was used with triple quadrupole and ion trap ESI-MS/MS systems for the identification of proteins at the low femtomole to subfemtomole sensitivity level [43, 44]. While this system is highly sensitive, its operation is labor-intensive and its operation has not been automated. In order to devise an analytical system with both the sensitivity of a CE and the level of automation of LC, we have constructed

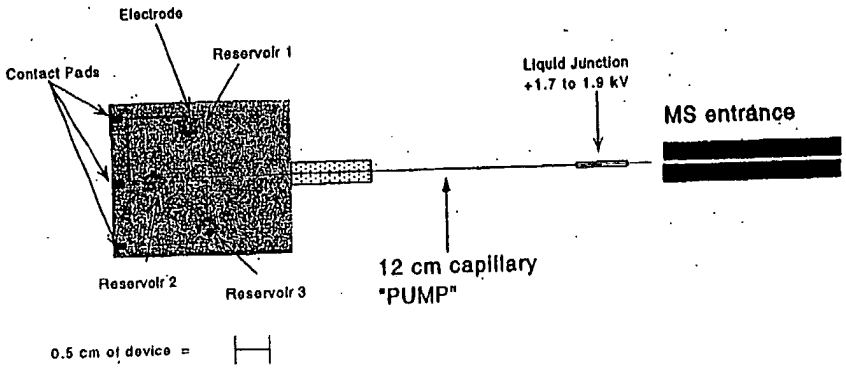


Figure 3. Schematic illustration of a microfabricated analytical system for CE, consisting of a micromachined device, coated capillary electroosmotic pump, and microelectrospray interface. The dimensions of the channels and reservoir are as indicated in the text. The channels on the device were graphically enhanced to make them more visible. Reproduced from [45], with permission.

microfabricated devices for the introduction of samples into ESI-MS for high-sensitivity peptide analysis.

The basic device is a piece of glass into which channels of 10–30 μm in depth and 50–70 μm in diameter are etched by using photolithography/etching techniques similar to the ones used in the semiconductor industry. (A simple device is shown in Fig. 3). The channels are connected to an external high voltage power supply [45]. Samples are manipulated on the device and off the device to the MS by applying different potentials to the reservoirs. This creates a solvent flow by electroosmotic pumping which can be redirected by changing the position of the electrode. Therefore, without the need for valves or gates and without any external pumping, the flow can be redirected by simply switching the position of the electrodes on the device. The direction and rate of the flow can be modulated by the size and the polarity of the electric field applied and also by the charge state of the surface.

The type of data generated by the system is illustrated in Fig. 4, which shows the mass spectrum of a peptide sample representing the tryptic digest of carbonic anhydrase at 290 fmol/μL. Each numbered peak indicates a peptide successfully identified as being derived from carbonic anhydrase. Some of the unassigned signals may be chemical or peptide contaminants. The MS is programmed to automatically select each peak and subject the peptide to CID. The resulting CID spectra are then used to identify the protein by correlation with sequence databases. Therefore, this system allows us to concurrently apply a number of protein digests onto the device, to sequentially mobilize the samples, to automatically generate CID spectra of selected peptide ions and to search sequence databases for protein identification. These steps are performed automatically without the need for user input and proteins can be identified at very low femtomole level sensitivity at a rate of approximately one protein per 15 min.

3.4 Assessment of 2-DE-MS proteome technology

Using a combination of the analytical techniques described above we have identified the 80 protein spots indicated in Fig. 5. The protein pattern was generated by separating a total of 40 microgram of protein contained in a total cell lysate of the yeast strain YPH499 by high resolution 2-DE and silver staining of the separated proteins. To estimate how far this type of proteome analysis can penetrate towards the identification of low abundance proteins, we have calculated the codon bias of the genes encoding the respective proteins. Codon bias is a

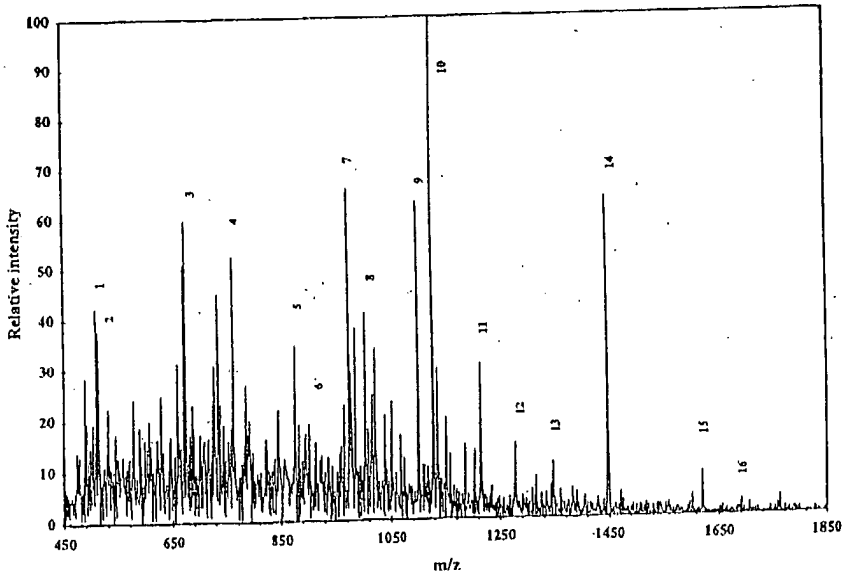


Figure 4. MS spectrum of a tryptic digest of carbonic anhydrase using the microfabricated system shown in Fig. 3. 290 fmol/μL of carbonic anhydrase tryptic digest was infused into a Finnigan LCQ ion trap MS. Each peak was selected for CID, and those which were identified as containing peptides derived from carbonic anhydrase are numbered. Reproduced from [45], with permission.

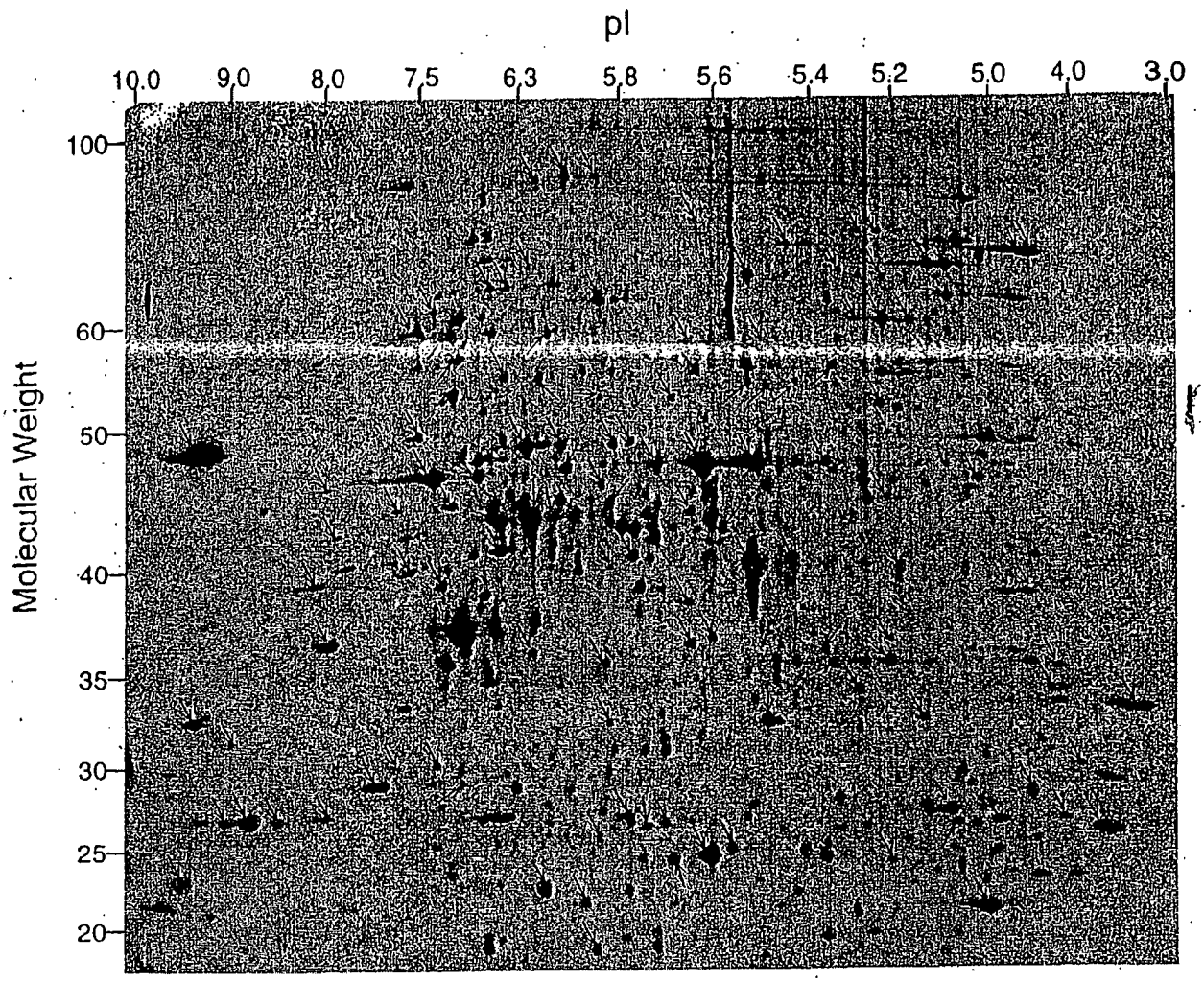


Figure 5. 2-DE separation of a lysate of yeast cells, with identified proteins highlighted. The first dimension of separation was an IPG from pH 3–10, and the second dimension was a 10%T SDS-PAGE gel. Proteins were visualized by silver staining. Further details of experimental procedures are included in S. P. Gygi *et al.* (submitted).

calculated measure of the degree of redundancy of triplet DNA codons used to produce each amino acid in a particular gene sequence. It has been shown to be a useful indicator of the level of the protein product of a particular gene sequence present in a cell [46]. The general rule which applies is that the higher the value of the codon bias calculated for a gene, the more abundant the protein product of that gene becomes. The calculated codon bias values corresponding to the proteins identified in Fig. 5 are shown in Fig. 6b. Nearly all of the proteins identified (> 95%) have codon bias values of > 0.2, indicating they are highly abundant in cells. In contrast, codon bias values calculated for the entire yeast genome (Fig. 6a) show that the majority of proteins present in the proteome have a codon bias of < 0.2 and are thus of low abundance.

This finding is of considerable importance in our assessment of the current status of proteome analysis technology. It is clear that even using highly sensitive analytical techniques, we are only able to visualize and identify the more abundant proteins. Since many important regulatory proteins are present only at low abundance, these would not be amenable to analysis using such techniques. This situation would be exacerbated in the analysis of proteomes containing many more proteins than the approximately 6000 gene products present in yeast cells [16]. In the analysis of, for example, the proteome of any human cells, there are potentially 50000–100000 gene products [47]. Inherent limitations on the amount of protein that can be loaded on 2-DE, and the number of components that can be resolved, indicate that only the most highly abundant fraction of the many gene products could be successfully analyzed. One approach that has been employed to circumvent these limitations is the use of very narrow range immobilized pH gradient strips for the first-dimension separation of 2-DE [48]. Since only those proteins which focus within the narrow range will enter the second dimension of separation, a much higher sample loading within the desired range is possible. This, in turn, can lead to the visualization and identification of less abundant proteins.

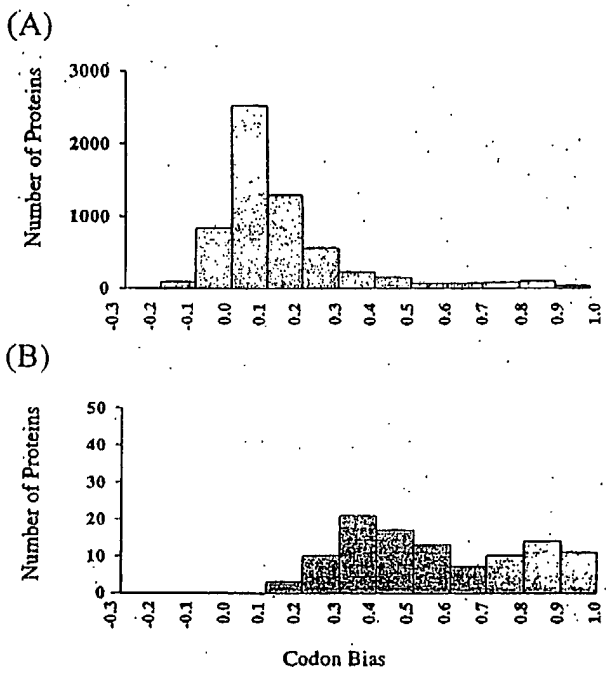


Figure 6. Calculated codon bias values for yeast proteins. (A) Distribution of calculated values for the entire yeast proteome. (B) Distribution of calculated values for the subset of 80 identified proteins also shown in Figs. 1 and 5. Further details of experimental procedures are included in S. P. Gygi *et al.* (submitted).

4 Utility of proteome analysis for biological research

For the success of proteomics as a mainstream approach to the analysis of biological systems it is essential to define how proteome analysis and biological research projects intersect. Without a clear plan for the implementation of proteome-type approaches into biological research projects the full impact of the technology can not be realized. The literature indicates that proteome analysis is used both as a database/data archive, and as a biological assay or biological research tool.

4.1 The proteome as a database

The use of proteomics as a database or data archive essentially entails an attempt to identify all the proteins in a cell or species and to annotate each protein with the known biological information that is relevant for each protein. The level of annotation can, of course, be extensive. The most common implementation of this idea is the separation of proteins by high resolution 2-DE, the identification of each detected protein spot and the annotation of the protein spots in a 2-DE gel database format. This approach is complicated by the fact that it is difficult to precisely define a proteome and to decide which proteome should be represented in the database. In contrast to the genome of a species, which is essentially static, the proteome is highly dynamic. Processes such as differentiation, cell activation and disease can all significantly change the proteome of a species. This is illustrated in Fig. 7. The figure shows two high-resolution 2-DE maps of proteins isolated from rat serum. Fig. 7A is from the serum of normal rats, while Fig. 7B is from the serum of rats in acute-phase serum after prior treatment with an inflammation-causing agent [49]. It is obvious that the protein patterns are significantly different in several areas, raising the question of exactly which proteome is being described.

Therefore, a comprehensive proteome database of a species or cell type needs to contain all of the parameters which describe the state and the type of the cells from which the proteins were extracted as well as the software tools to search the database with queries which reflect the dynamics of biological systems. A comprehensive proteome database should be capable of quantitatively describing the fate of each protein if specific systems and pathways are activated in the cell. Specifically, the quantity, the degree of modification, the subcellular location and the nature of molecules specifically interacting with a protein as well as the rate of change of these variables should be described. Using these admittedly stringent criteria, there is currently no comlete proteome database. A number of such databases are, however, in the process of being constructed. The most advanced among them, in our opinion, are the yeast protein database YPD [50] (accessible at http://www.ypd.com) and the human 2D-PAGE databases of the Danish Centre for Human Genome Research [12] (accessible at http://biobase.dk/cgi-bin/celis). While neither can be considered complete as not all of the potential gene products are identified, both contain extensive annotation of supplemental information for many of the spots which are positively identified in reference samples.

4.2 The proteome as a biological assay

The use of proteome analysis as a biological assay or research tool represents an alternative approach to integrating biology with proteomics. To investigate the state of a system, samples are subjected to a specific proceess that allows the quantitative or qualitative measurement of some of the variables which describe the system. In typical biochemical assays one variable (*e.g.*, enzyme activity) of a single component (*e.g.*, a particular enzyme) is measured. Using proteomics as an assay, multiple variables (*e.g.*, expression level, rate of synthesis, phosphorylation state, etc.) are measured concurrently on many (ideally all) of the proteins in a sample. The use of proteomics as an assay is a less far-reaching proposition than the construction of a comprehensive proteome database. It does, however, represent a pragmatic approach which can be adapted to investigate specific systems and pathways, as long as the interpretation of the results takes into account that with current technology not all of the variables which describe the system can be observed (see Section 3.4).

A common implementation of proteome analysis as a biological assay is when a 2-DE protein pattern generated from the analysis of an experimental sample is compared to an array of reference patterns representing different states of the system under investigation. The state of the experimental system at the time the sample was generated is therefore determined by the quantita-

tive comparative analysis of hundreds to a few thousand proteins. Comparative analysis of the 2-DE patterns furthermore highlights quantitative and qualitative differences in the protein profiles which correlate with the state of the system. For this type of analysis it is not essential that all the proteins are identified or even visualized, although the results become more informative as more proteins are compared. It is obvious, however, that the possibility to identify any protein deemed characteristic for a particular state dramatically enhances this approach by opening up new avenues for experimentation.

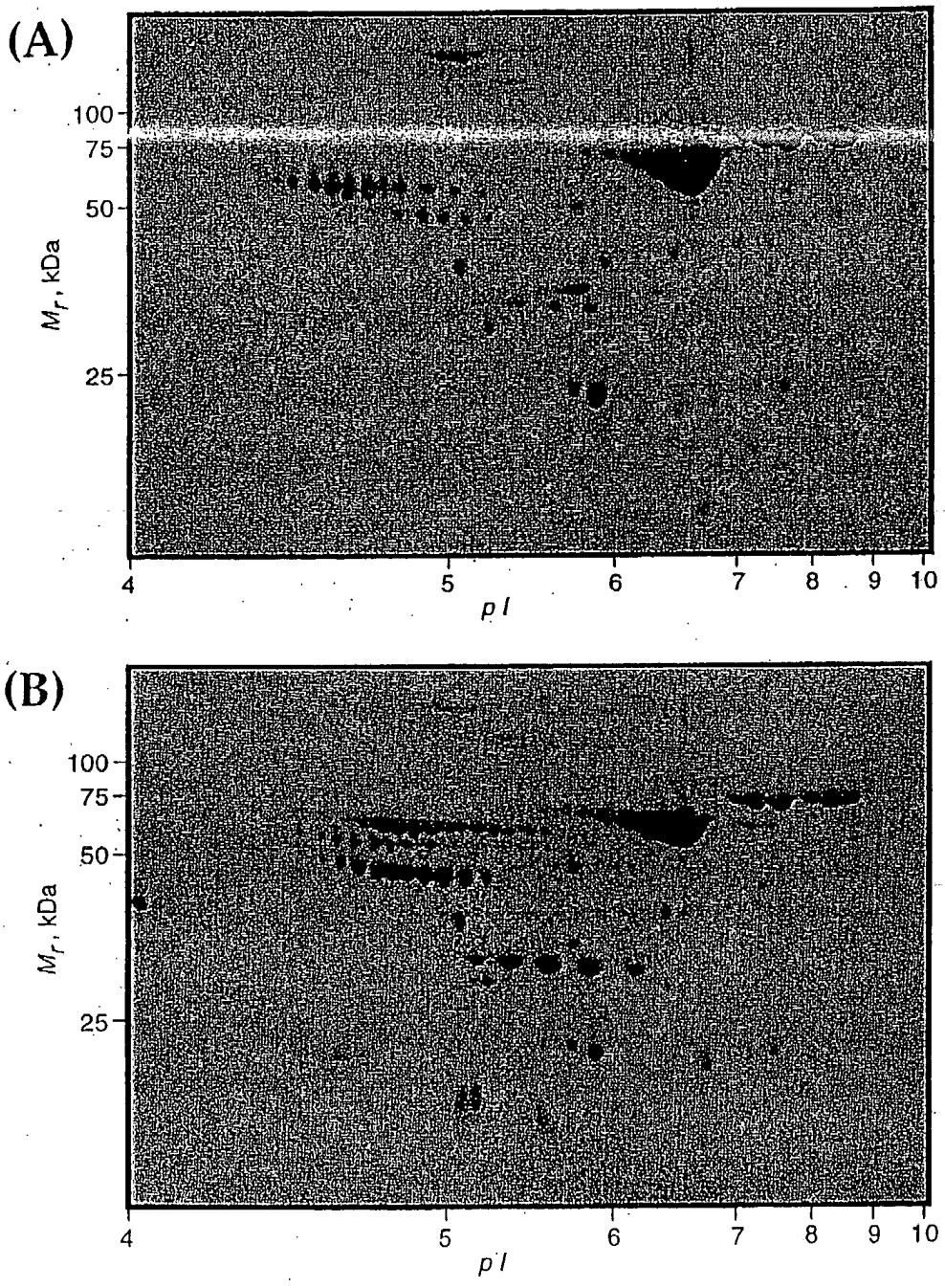


Figure 7. High resolution 2-DE map of proteins isolated from rat serum with or without prior exposure to an inflammation-causing agent. (A) normal rat serum, (B) acute-phase serum from rats which had previously been exposed to an inflammation-causing agent. The first dimension of separation is an IPG from pH 4–10, and the second dimension is a 7.5–17.5%T gradient SDS-PAGE gel. Proteins were visualized by staining with amido black. Further details of experimental procedures are included in [14, 49].

Proteome analysis as a biological assay has been successfully used in the field of toxicology, to characterize disease states or to study differential activation of cells. The approach is limited, of course, by the fact that only the visible protein spots are included in the assay, and it is well known that a substantial but far from complete fraction of cellular proteins are detected if a total cell lysate is separated by 2-DE. Proteins may not be detected in 2-DE gels because they are not abundant enough to be visualized by the detection method used, because they do not migrate within the boundaries (size, p*I*) resolved by the gel, because they are not soluble under the conditions used, or for other reasons.

A different way to use proteome analysis as a biological assay to define the state of a biological system is to take advantage of the wealth of information contained in 2-DE protein patterns. 2-DE is referred to as two-dimensional because of the electrophoretic mobility and the isoelectric points which define the position of each protein in a 2-DE pattern. In addition to the two dimensions used to generate the protein patterns, a number of additional data dimensions are contained in the protein patterns. Some of these dimensions such as protein expression level, phosphorylation state, subcellular location, association with other proteins, rate of synthesis or degradation indicate the activity state of a protein or a biological system. Comparative analysis of 2-DE protein patterns representing different states is therefore ideally suited for the detection, identification and analysis of suitable markers. Once again it must be emphasized that in this type of experiment only a fraction of the cellular proteins is analyzed. Since many regulatory proteins are of low abundance, this limitation is a concern, particularly in cases in which regulatory pathways are being investigated.

5 Concluding remarks

In this report we have addressed three main issues related to proteome analysis. First, we have discussed the rationale for studying proteomes. Second, we have assessed the technical feasibility of analyzing proteomes and described current proteome technology, and third, we have analyzed the utility of proteome analysis for biological research. It is apparent that proteome analysis is an essential tool in the analysis of biological systems. The multi-level control of protein synthesis and degradation in cells means that only the direct analysis of mature protein products can reveal their correct identities, their relevant state of modification and/or association and their amounts. Recently developed methods have enabled the identification of proteins at ever-increasing sensitivity levels and at a high level of automation of the analytical processes. A number of technical challenges, however, remain. While it is currently possible to identify essentially any protein spots that can be visualized by common staining methods, it is apparent that without prior enrichment only a relatively small and highly selected population of long-lived, highly expressed proteins is observed. There are many more proteins in a given cell which are not visualized by such methods. Frequently it is the low abundance proteins that execute key regulatory functions.

We have outlined the two principal ways proteome analysis is currently being used to intersect with biological research projects: the proteome as a database or data archive and proteome analysis as a biological assay. Both approaches have in common that at present they are conceptually and technically limited. Current proteome databases typically are limited to one cell type and one state of a cell and therefore do not account for the dynamics of biological systems. The use of proteome analysis as a biological assay can provide a wealth of information, but it is limited to the proteins detected and is therefore not truly proteome-wide. These limitations in proteomics are to a large extent a reflection of the fact that proteins in their fully processed form cannot easily be amplified and are therefore difficult to isolate in amounts sufficient for analysis or experimentation. The fact that to date no complete proteome has been described further attests to these difficulties. With continued rapid progress in protein analysis technology, however, we anticipate that the goal of complete proteome analysis will eventually become attainable.

We would like to acknowledge the funding for our work from the National Science Foundation Science and Technology Center for Molecular Biotechnology and from the NIH. We thank Yvan Rochon and Bob Franza for providing the yeast gel shown and Elisabetta Gianazza for providing the rat serum gels shown.

Received April 21, 1998

6 References

[1] Wilkins, M. R., Pasquali, C., Appel, R. D., Ou, K., Golaz, O., Sanchez, J.-C., Yan, J. X., Gooley, A. A., Hughes, G., Humphery-Smith, I., Williams, K. L., Hochstrasser, D. F., *Bio/Technology* 1996, *14*, 61–65.

[2] Hodges, P. E., Payne, W. E., Garrels, J. I., *Nucleic Acids Res.* 1998, *26*, 68–72.

[3] O'Connor, C. D., Farris, M., Fowler, R., Qi, S. Y., *Electrophoresis* 1997, *18*, 1483–1490.

[4] Cordwell, S. J., Basseal, D. J., Humphery-Smith, I., *Electrophoresis* 1997, *18*, 1335–1346.

[5] Urquhart, B. L., Atsalos, T. E., Roach, D., Basseal, D. J., Bjellqvist, B., Britton, W. L., Humphery-Smith, I., *Electrophoresis* 1997, *18*, 1384–1392.

[6] Wasinger, V. C., Bjellqvist, B., Humphery-Smith, I., *Electrophoresis* 1997, *18*, 1373–1383.

[7] Link, A. J., Hays, L. G., Carmack, E. B., Yates III, J. R., *Electrophoresis* 1997, *18*, 1314–1334.

[8] Sazuka, T., Ohara, O., *Electrophoresis* 1997, *18*, 1252–1258.

[9] VanBogelen, R. A., Abshire, K. Z., Moldover, B., Olson, E. R., Neidhardt, F. C., *Electrophoresis* 1997, *18*, 1243–1251.

[10] Guerreiro, N., Redmond, J. W., Rolfe, B. G., Djordjevic, M. A., *Mol. Plant Microbe Interact.* 1997, *10*, 506–516.

[11] Yan, J. X., Tonella, L., Sanchez, J.-C., Wilkins, M. R., Packer, N. H., Gooley, A. A., Hochstrasser, D. F., Williams, K. L., *Electrophoresis* 1997, *18*, 491–497.

[12] Celis, J., Gromov, P., Ostergaard M., Madsen, P., Honoré, B., Dejgaard, K., Olsen, E., Vorum, H., Kristensen, D. B., Gromova, I., Haunso, A., Van Damme, J., Puype, M., Vandekerckhove, J., Rasmussen, H. H., *FEBS Lett.* 1996, *398*, 129–134.

[13] Appel, R. D., Sanchez, J.-C., Bairoch, A., Golaz, O., Miu, M., Vargas, J. R., Hochstrasser, D. F., *Electrophoresis* 1993, *14*, 1232–1238.

[14] Haynes, P., Miller, I., Aebersold, R., Gemeiner, M., Eberini, I., Lovati, R. M., Manzoni, C., Vignati, M., Gianazza, E., *Electrophoresis* 1998, *19*, 1484–1492.

[15] Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J.-F., Dougherty, B. A., Merrick, J. M., McKenney, K., Sutton, G., FitzHugh, W., Fields, C., Gocayne, J. D., Scott, J., Shirley, R., Liu, L.-I., Glodek, A., Kelley, J. M., Weidman, J. P., Phillips, C. A., Spriggs, T., Hedblom, E., Cotton, M. D., Utterback, T. R., Hanna, N. C., Nguyen, D. T., Saudek, D. M., Brandon, R. C., Fine, L. D., Fritchman, J. L., Fuhrmann, J. L., Geoghagen, N. S. M., Gnehm, C. L., McDonald, L. A., Small, K. V., Fraser, C. M., Smith, C. O., Venter, J. C., *Science* 1995, *269*, 496–512.

[16] Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H., Oliver, S. G., *Science* 1996, *274*, 546.

[17] Fraser, C. M., Casjens, S., Huang, W. M., Sutton, G. G., Clayton, R., Lathigra, R., White, O., Ketchum, K. A., Dodson, R., Hickey, E. K., Gwinn, M., Dougherty, B., Tomb, J. F., Fleischmann, R. D., Richardson, D., Peterson, J., Kerlavage, A. R., Quackenbush, J., Salzberg, S., Hanson, M., van Vugt, R., Palmer, N., Adams, M. D., Gocayne, J., Weidman, J., Utterback, T., Watthey, T., McDonald, L., Artiach, P., Bowman, C., Garland, S., Fujii, C., Cotton, M. D., Horst, K., Roberts, K., Hatch, B., Smith, H. O., Venter, J. C., *Nature* 1997, *390*, 580–586.

[18] Liang, P., Pardee, A. B., *Science* 1992, *257*, 967–971.

[19] Lashkari, D. A., DeRisi, J. L., McCusker, J. H., Namath, A. F., Gentile, C., Hwang, S. Y., Brown, P. O., Davis, R. W., *Proc. Natl. Acad. Sci. USA* 1997, *94*, 13057–13062.

[20] Shalon, D., Smith, S. J., Brown, P. O., *Genome Res.* 1996, *6*, 639–645.

[21] Velculescu, V. E., Zhang, L., Vogelstein, B., Kinzler, K. W., *Science* 1995, *270*, 484–487.

[22] Velculescu, V. E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M. A., Bassett, D. E., Hieter, P., Vogelstein, B., Kinzler, K. W., *Cell* 1997, *88*, 243–251.

[23] Krishna, R. G., Wold, F., *Adv. Enzymol.* 1993, *67*, 265–298.

[24] Görg, A., Postel, W., Gunther, S., *Electrophoresis* 1988, *9*, 531–546.

[25] Klose, J., Kobalz, U., *Electrophoresis* 1995, *16*, 1034–1059.

[26] Matsudaira, P., *J. Biol. Chem.* 1987, *262*, 10035–10038.

[27] Aebersold, R. H., Teplow, D. B., Hood, L. E., Kent, S. B., *J. Biol. Chem.* 1986, *261*, 4229–4238.

[28] Rosenfeld, J., Capdevielle, J., Guillemot, J. C., Ferrara, P., *Anal. Biochem.* 1992, *203*, 173–179.

[29] Aebersold, R. H., Leavitt, J., Saavedra, R. A., Hood, L. E., Kent, S. B., *Proc. Natl. Acad. Sci. USA* 1987, *84*, 6970–6974.

[30] Honoré, B., Leffers, H., Madsen, P., Celis, J. E., *Eur. J. Biochem.* 1993, *218*, 421–430.

[31] Mann, M., Wilm, M., *Anal. Chem.* 1994, *66*, 4390–4399.

[32] Eng, J., McCormack, A. L., Yates III, J. R., *J. Amer. Mass Spectrom.* 1994, *5*, 976–989.

[33] Yates III, J. R., Eng, J. K., McCormack, A. L., Schieltz, D., *Anal. Chem.* 1995, *67*, 1426–1436.

[34] Shevchenko, A., Wilm, M., Vorm, O., Mann, M., *Anal. Chem.* 1996, *68*, 850–858.

[35] Hess, D., Covey, T. C., Winz, R., Brownsey, R. W., Aebersold, R., *Protein Sci.* 1993, *2*, 1342–1351.

[36] van Oostveen, I., Ducret, A., Aebersold, R., *Anal. Biochem.* 1997, *247*, 310–318.

[37] Lui, M., Tempst, P., Erdjument-Bromage, H., *Anal. Biochem.* 1996, *241*, 156–166.

[38] Patterson, S. D., Aebersold, R. A., *Electrophoresis* 1995, *16*, 1791–1814.

[39] Ducret, A., Foyn, Brunn, C., Bures, E. J., Marhaug, G., Husby, G. R. A., *Electrophoresis* 1996, *17*, 866–876.

[40] Haynes, P. A., Fripp, N., Aebersold, R., *Electrophoresis* 1998, *19*, 939–945.

[41] Figeys, D., Van Oostveen, I., Ducret, A., Aebersold, R., *Anal. Chem.* 1996, *68*, 1822–1828.

[42] Ducret, A., Van Oostveen, I., Eng, J. K., Yates III, J. R., Aebersold, R., *Protein Sci.* 1997, *7*, 706–719.

[43] Figeys, D., Ducret, A., Yates III, J. R., Aebersold, R., *Nature Biotech.* 1996, *14*, 1579–1583.

[44] Figeys, D., Aebersold, R., *Electrophoresis* 1997, *18*, 360–368.

[45] Figeys, D., Ning, Y., Aebersold, R., *Anal. Chem.* 1997, *69*, 3153–3160.

[46] Garrels, J. I., McLaughlin, C. S., Warner, J. R., Futcher, B., Latter, G. I., Kobayashi, R., Schwender, B., Volpe, T., Anderson, D. S., Mesquita-Fuentes, R., Payne, W. E., *Electrophoresis* 1997, *18*, 1347–1360.

[47] Schuler, G. D., Boguski, M. S., Stewart, E. A., Stein, L. D., Gyapay, G., Rice, K., White, R. E., Rodriguez-Tome, P., Aggarwal, A., Bajorek, E., Bentolila, S., Birren, B. B., Butler, A., Castle, A. B., Chiannilkulchai, N., Chu, A., Clee, C., Cowles, S., Day, P. J., Dibling, T., Drouot, N., Dunham, I., Duprat, S., Edwards, C., Fan, J.-B., Fang, N., Fizames, C., Garrett, C., Green, L., Hadley, D., Harris, M., Harrison, P., Brady, S., Hicks, A., Holloway, E., Hui, L., Hussain, S., Louis-Dit-Sully, C., Ma, J., MacGilvery, A., Mader, C., Maratukulam, A., Matise, T. C., McKusick, K. B., Morissette, J., Mungall, A., Muselet, D., Nusbaum, H. C., Page, D. C., Peck, A., Perkins, S., Piercy, M., Qin, F., Quackenbush, J., Ranby, S., Reif, T., Rozen, S., Sanders, X., She, X., Silva, J., Slonim, D. K., Soderlund, C., Sun, W.-L., Tabar, P., Thangarajah, T., Vega-Czarny, N., Vollrath, D., Voyticky, S., Wilmer, T., Wu, X., Adams, M. D., Auffray, C., Walter, N. A. R., Brandon, R., Dehejia, A., Goodfellow, P. N., Houlgatte, R., Hudson, J. R., Jr., Ide, S. E., Iorio, K. R., Lee, W. Y., Seki, N., Nagase, T., Ishikawa, K., Nomura, N., Phillips, C., Polymeropoulos, M. H., Sandusky, M., Schmitt, K., Berry, R., Swanson, K., Torres, R., Venter, J. C., Sikela, J. M., Beckmann, J. S., Weissenbach, J., Myers, R. M., Cox, D. R., James, M. R., Bentley, D., *et al. Science* 1996, *274*, 540–546.

[48] Sanchez, J.-C., Rouge, V., Pisteur, M., Ravier, F., Tonella, L., Moosmayer, M., Wilkins, M. R., Hochstrasser, D. F., *Electrophoresis* 1997, *18*, 324–327.

[49] Miller, I., Haynes, P., Gemeiner, M., Aebersold, R., Manzoni, C., Lovati, M. R., Vignati, M., Eberini, I., Gianazza, E., *Electrophoresis* 1998, *19*, 1493–1500.

[50] Garrels, J. I., *Nucleic Acids Res.* 1996, *24*, 46–49.